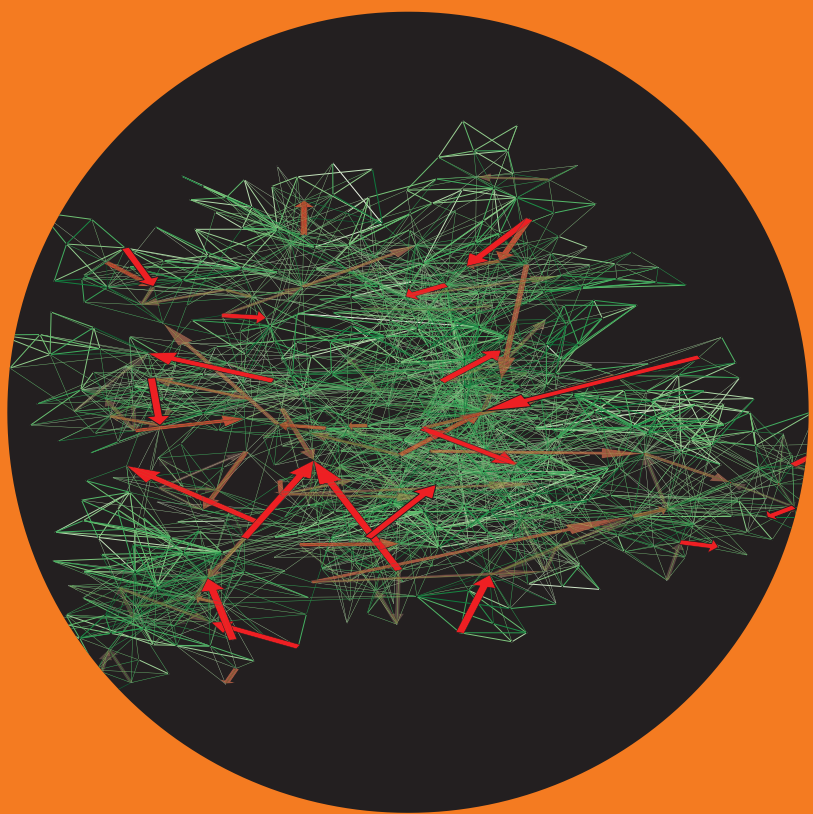
Department of Biomedical Engineering and Computational Science

Weighted and Temporal Networks: Towards More Realistic Representations of Complex Systems

Mikko Kivelä



A!

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 103/2012

WEIGHTED AND TEMPORAL NETWORKS: TOWARDS MORE REALISTIC REPRESENTATIONS OF COMPLEX SYSTEMS

Mikko Kivelä

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall F239a of the school on 5 October 2012 at 12.

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science

Supervising professor
Kimmo Kaski

Thesis advisor
Jari Saramäki

Preliminary examiners
Renaud Lambiotte
University of Namur
Belgium

Gergely Palla
Eötvös University
Hungary

Opponents
Petter Holme
Umeå University
Sweden

Aalto University publication series
DOCTORAL DISSERTATIONS 103/2012

ISBN 978-952-60-4758-4 (printed)
ISBN 978-952-60-4759-1 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-4759-1

Unigrafia Oy
Helsinki 2012

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Mikko Kivelä

Name of the doctoral dissertation
WEIGHTED AND TEMPORAL NETWORKS: TOWARDS MORE REALISTIC REPRESENTATIONS OF COMPLEX SYSTEMS

Publisher School of Science

Unit Department of Biomedical Engineering and Computational Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 103/2012

Field of research Complex networks

Manuscript submitted 5 March 2012 **Date of the defence** 5 October 2012

Permission to publish granted (date) 11 May 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Complex systems consisting of large numbers of interacting elements often display emergent behavior, which cannot be understood by the reductionistic approach of describing the elements and interactions in detail and in isolation. The complex networks framework takes a completely opposite approach by describing the elements and interactions as simply as possible focusing on the system-level behavior instead. This approach has been successful in identifying basic structural properties of systems from various fields such as biology, sociology, neuroscience, and technology. However, this very simplicity of description that makes the complex networks approach so versatile is also its main stumbling block. This is because for many systems, details of individual interactions such as their strengths or timings are essential. Some of such details can be taken into account with weighted and temporal networks. These ways of looking at networks are still fairly underdeveloped; however, there is currently intensive research, especially on temporal networks.

The contribution of this Thesis can be divided to three parts: First, it deepens the understanding of many multiscale phenomena in complex networks such as community structure, percolation, and social network formation. Second, it expands the borders of weighted network analysis, amongst others by pointing out the problem that many existing methods for studying cluster structure in weighted networks are domain-specific, instead of being generally applicable to all systems. It also introduces an improvement to clique percolation, which is a non-parametric method for finding cluster structure in weighted networks, that makes it more powerful both computationally and in describing the clusters. Third, it introduces new concepts and methods related to the emerging field of temporal networks and dynamics on top of them. It introduces a systematic way of using reference models to study temporal networks. Using this framework together with a temporal network of mobile communication, it is shown that bursty interaction sequences can slow down dynamics on top of temporal networks, and that such temporal effects can be as important as the network topology.

Keywords Complex networks, weighted networks, temporal networks, complex systems

ISBN (printed) 978-952-60-4758-4 **ISBN (pdf)** 978-952-60-4759-1

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 201 **urn** http://urn.fi/URN:ISBN:978-952-60-4759-1

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Mikko Kivelä

Väitöskirjan nimi
Painotetut ja aikariippuvaiset verkostot: kohti realistisempia kompleksisten systeemien esitysmuotoja

Julkaisija Perustieteiden korkeakoulu

Yksikkö Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 103/2012

Tutkimusala Verkostoteoria

Käsikirjoituksen pvm 05.03.2012 **Väitöspäivä** 05.10.2012

Julkaisuluvan myöntämispäivä 11.05.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Suuren määrän keskenään monimutkaisessa vuorovaikutuksessa olevia osasia sisältävät systeemit synnyttävät usein emergenttejä ilmiöitä, joita ei voida ymmärtää reduktionistisella lähestymistavalla, jossa nämä osaset ja vuorovaikutukset pyritään kuvaamaan mahdollisimman tarkasti ja toisistaan riippumatta. Kompleksisten verkostojen tutkimus perustuu täysin päinvastaiseen lähestymistapaan, jossa systeemin osaset ja vuorovaikutukset kuvataan mahdollisimman yksinkertaisesti ja keskitytään sen sijaan koko systeemin käytöksen kuvaamiseen. Tällä lähestymistavalla on menestyksekkäästi kuvailtu useisiin eri aloihin, kuten biologiaan, sosiologiaan, neurotieteeseen ja teknologiaan, liittyvien systeemien rakennetta. Toisaalta, samainen verkostojen käytön monimuotoisuuden mahdollistanut osasten ja vuorovaikutusten yksinkertainen kuvaustapa on myös tämän menetelmän suurin kompastuskivi. Tämä johtuu siitä, että yksittäisten vuorovaikutusten tarkemmat ominaisuudet, kuten niiden voimakkuus tai ajoitus, ovat olennainen osa monia systeemejä. Osaa näistä ominaisuuksista voidaan tutkia painotettujen ja aikariippuvaisten verkostojen avulla. Nämä näkökulmat verkostoihin ovat vielä alikehittyneet, vaikkakin erityisesti aikariippuvaiset verkostot ovat tällä hetkellä intensiivisen tutkimuksen kohteena.

Tämän väitöskirjan anti verkostotutkimukselle voidaan jakaa kolmeen osaan: Ensinnäkin, se syventää verkostoissa monilla eri rakenteen tasoilla tapahtuvien ilmiöiden ymmärrystä mm. yhteisörakenteen, perkolaation ja sosiaalisten verkostojen syntymekanismien osalta. Toisekseen, se laajentaa painotettujen verkostojen teoriaa mm. osoittamalla monien painotettujen klusterointimenetelmien olevan käyttökelpoisia vain tietyillä tutkimusalueilla sen sijaan, että ne olisivat yleiskäyttöisiä. Lisäksi väitöskirjassa parannellaan klikkiperkolaatiomenetelmää, joka on eräs yhteisörakenteen etsintämenetelmä, siten että siitä tulee yhteisörakennetta paremmin kuvaava ja laskennallisesti vähemmän vaativa. Kolmanneksi, tässä työssä esitellään uusia käsitteitä ja menetelmiä aikariippuvaisille verkostoille ja niiden päällä toimiviin dynaamisiin prosesseihin liittyen. Tässä työssä esitellään myös systemaattinen tapa käyttää referenssimalleja aikariippuvaisten verkostojen tutkimiseen. Tätä tapaa ja matkapuhelinverkossa tapahtuvaa kommunikaatiota esimerkkinä käyttäen näytetään kuinka ajassa purskeisiin jakautuvat yhteydenotot voivat hidastaa dynaamisia prosesseja aikariippuvaisissa verkostoissa, ja kuinka aikaulottuvuuden mukaanotto saattaa olla yhtä tärkeää kuin verkoston topologian huomiointi.

Avainsanat verkostoteoria, painotetut verkot, aikariippuvaiset verkot, kompleksiset systeemit

ISBN (painettu) 978-952-60-4758-4 **ISBN (pdf)** 978-952-60-4759-1

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2012

Sivumäärä 201 **urn** http://urn.fi/URN:ISBN:978-952-60-4759-1

Preface

This thesis has been prepared in the Department of Biomedical Engineering and Computation Science (former Laboratory of Computational Engineering) at Aalto University (former Helsinki University of Technology), and it concludes my Doctor of Technology degree. I have worked for the degree during the years 2009-2012, but some of the work was done already when I was a research student (2006-2008) and wrote my masters thesis (2008-2009).

All of my work presented in this thesis was completed as a member of the complex networks research group at BECS(LCE). My foremost gratitude goes to my thesis instructor and the leader of the research group Prof. Jari Saramäki for introducing me to the field of complex networks and providing insight and guidance throughout the years I worked in the lab. I'm also thankful for my thesis supervisor Prof. Kimmo Kaski for making the lab such a great place to do research.

I'm grateful for being able to participate in so many interesting research projects with so many smart people. My special thanks go to all of my coauthors I worked with (who are not mentioned before): Prof. Jukka-Pekka Onnela, Prof. János Kertész, Dr. Jussi Kumpula, Dr. Riitta Toivonen, Lauri Kovanen, Dr. Andrea Lancichinetti, Prof. Santo Fortunato, Dr. Márton Karsai, Dr. Raj Kumar Pan, and Mikko Viinikainen. In addition, I was lucky to have such a good coworkers as Dr. Riku Linna, Jörkki Hyvönen, Dr. Tapio Heimo, Jenni Hulkkonen, Gerardo Iñiguez, Dr. Hang-Hyun Jo, Dr. Vasyl Palchykov, Markus Karppinen, and Katri Kaunismaa, who I had countless useful discussions in the office and even more, maybe not so useful but definitely entertaining, discussions during lunch and coffee.

Finally, I would like to thank my family and friends for all the support and encouragement throughout my studies.

Espoo, August 28, 2012,

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski and J. Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev. E*, 75, 027105, 2007.

II J.M. Kumpula, M. Kivelä, K. Kaski,and J. Saramäki. A sequential algorithm for fast clique percolation. *Phys. Rev. E*, 78, 026109, 2008.

III R. Toivonen, L. Kovanen, M. Kivelä, J.-P. Onnela, J. Saramäki, and K. Kaski. A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks*, Vol. 31, Issue 4, 240–254, 2009.

IV A. Lancichinetti, M. Kivelä, J. Saramäki, S. Fortunato. Characterizing the Community Structure of Complex Networks. *PLoS ONE*, 5(8): e11976, 2010.

V M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, J. Saramäki. Small But Slow World: How Network Topology and Burstiness Slow Down Spreading. *Phys. Rev. E*, 83, 025102, 2011.

VI R. K. Pan, M. Kivelä, J. Saramäki, K. Kaski, J. Kertész. Using explosive percolation in analysis of real-world networks. *Phys. Rev. E*, 83, 046112, 2011.

VII R. Toivonen, M. Kivelä, J. Saramäki, M. Viinikainen, M. Vanhatalo, M. Sams. Networks of Emotion Concepts. *PLoS ONE*, 7(1): e28883, 2012.

VIII M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, M. Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, P03005, 2012.

Author's contribution

The research reported in this thesis is a result of collaboration between the author of this thesis, Mikko Kivelä, and the other authors of the included publications. Kivelä was the principal author in Publication VIII. He was the main contributor to the theoretical analysis and preparation of the manuscript, and had main responsibility in analysis of the numerical results. As a second author in Publications I, II, IV, V, VI, and VII, and as a third author in Publication III, Kivelä participated in preparation of the manuscripts, and contributed significantly in developing the research and ideas. In Publications IV, V, VI, and VII he was one of the main contributors in developing computer programs and analysing the data. In Publication II he contributed by developing and analysing the algorithms. In Publication I he developed the computer programs and used them to produce the numerical results, and derived all analytical results. Further, in Publications V and VI, Kivelä designed and performed the initial, proof-of-concept numerical analysis that triggered these research projects.

List of Figures

1. Introduction

Nowadays it is not uncommon to write one's thesis about a narrow subject in some very specialized field of science. In medicine, one might consider the effects of single gene to a disease, or in chemistry, one might be interested in the synthesis of a single compound. This kind of specialization is characteristic to reductionistic science; a way of studying systems by reducing them to smaller and smaller parts. This paradigm has dominated the science of the 20th century and brought us many good things from quantum mechanics to the discovery of DNA as basis of genetic information. However, it is becoming increasingly clear that the reductionistic approach has its limits: one cannot unravel how genetic regulation works (or doesn't work, in the case of cancer) by studying a single gene, understand how consciousness is born in the brain by observing a single neuron, or predict changes in the society by analyzing the mind-set of a single person in isolation.

This thesis studies scientific collaboration, mobile phone call patterns, interactions of proteins, global trading, online auctions, and organization of emotions, among other things. Such diversity in research topics is typical for the science of complex systems. Here the paradigm is the opposite to that of reductionism: instead of zooming into the individual parts of the system, we try to describe the parts as simply as possible, and instead zoom out and focus on how the interactions between the parts are structured on the system-wide scale. Remarkably, in most systems with non-trivial interactions, emergent behavior is observed: patterns emerge that would be hard to predict from the behavior of individual agents or interactions.

Complex systems started to become popular around the 80's, mostly among statistical physicists [1]. The system-level approach was very natural for physicists since they had already found out that the reductionistic

approach, such as the one taken by quantum mechanics, is not enough to explain all the physics [2]. Instead, many physical phenomena, such as thermodynamics, can be understood by simplifying the model for the particles and considering how they work as system. Today the study of complex systems is a highly interdisciplinary field, which is mostly due to an explosion in the amount of data available to scientists in many fields such as sociology, genetics, epidemiology, and neuroscience. Also the growth in both number and power of computers has made it possible to perform increasingly complicated analyses of this new data.

This thesis concentrates on complex networks, which is currently the most popular approach for studing complex systems. This approach is popular since many systems are composed of some agents that interact with each other, and much of the complexity in these systems can be traced back to the structure of the network formed by these interactions. The field of complex networks was born at late 90's when it was found out that regular lattices or random networks were not good enough models for empirical systems [3, 4]. Instead, real-world networks contain nontrivial structure that is interesting in itself, but also has a large impact on processes taking place on top of the networks, and on the set of possible mechanisms that could have been responsible for the evolution of the networks. Since then, the use of network theory in science has grown dramatically, with thousands of complex networks related publications appearing per year.

By far, most of the work in complex networks so far has been devoted to static binary networks. Here, "binary" means that each pair of nodes in the network is either connected or not. As an example, in a social network this would mean that you either know someone or not, and all other information about the relationship is irrelevant. However, it is known that tie strengths (or link weights [5]) are very important in the context of social networks [6]. A network is static if it doesn't change in time. This might be a valid assumption for example when finding weak spots in a power grid [7], but again in social networks and for among others in airline traffic networks, information on the dynamics and timings of interactions can be essential [8–11]. In this thesis, the main theme is to take a small step back towards reductionism by including more detailed information on link weights and timings in the complex networks framework.

The thesis is organized as follows: First, the main concepts of network theory are introduced in Chapter II. This Chapter also discusses some

of the main results of the field. Chapter III concentrates on taking the weights of the links into account and deals with the inherent problems that any attempts to generalize unweighted network concepts are faced with. Chapter IV introduces temporal networks and goes through related concepts and some results. Finally, Chapter V reviews the results in all the Publications of this thesis and discusses future aspects.

2. Complex networks

The basis of complex networks theory is the notion that the essence of most complex systems can be captured by representing them as networks, which consist of the elementary parts of the system (nodes) and their pairwise interactions (links). Everything else can be left out. The simplicity of the network abstraction is the reason why so many different systems can be mapped to this unified framework: In a social system people are linked through relationships, the routers on the Internet are connected by wires, and in a cell the genes interact by controlling each others' expression. The complexity of all of these systems is due to the non-trivial topology of the connections, and hence, all of them can be studied with the tools of complex networks analysis. Remarkably, the network representations of most of these complex systems display similar characteristics to such a degree that one might even describe them as universal properties of complex networks.

In mathematics, networks have long been studied under the name of graphs. The first paper in graph theory is considered to be the one published in 1736 by Euler. In this paper, the nodes of the graph are four parts of the city of Königsberg, separated by rivers running across the city, and the links are the seven bridges which connected these parts at the time. Later, graph theory became an important part of discrete mathematics. However, mathematicians were mainly concerned on constructing graphs with simple deterministic rules or about very general results that could be derived for all possible graphs. Outside of mathematics, graphs were also used to describe some small empirical systems, such as social networks, or sociograms, introduced by Moreno in the 1930's [12]. Clearly, the lack of computers for analyzing data and automation in data collection made it impossible to consider any large empirical systems as networks.

A big theoretical step towards rigorous analysis of large-scale complex

systems using networks was taken by Erdős and Rényi [13] in the late 1950's. They studied the properties of random graphs, where some fixed number of edges, L, is distributed uniformly at random between all possible node pairs. Even though this basic Erdős and Rényi (ER) model is simple, on a large scale the random graphs display emergent properties which have remained an active research topic in mathematics long after the publication of the model [14]. As model of empirical networks the random graphs were adopted as a part of social network analysis [15, 16] and they are still one of the central concepts in the complex networks literature.

In the late 90's, two landmark papers were published that showed that neither regular lattice models or random graphs were enough to explain the topology of many empirical systems. Instead, many networks were shown to be small worlds [3] and scale-free [4]. These findings triggered a large interest especially in the statistical physics community, leading to the field of complex networks as we now know it.

In this Chapter, we will briefly go through some of the basic concepts of network theory, such as small-world and scale-free networks. We will also discuss other important related topics such as percolation, community structure, and social network models. The purpose of this Chapter and the whole introductory part of this Thesis is not to give a full review on complex networks, or even to be a short introduction to the whole subject, but to provide the reader with the necessary concepts and background in order to follow the rest of the thesis. Thus, some central concepts such as centrality measures and dynamics on top of networks will be left out. However, we will discuss spreading dynamics on networks in Chapter III, where temporal networks are discussed. There are several text books [17–21] and review articles about complex networks [22–25], and even some popular science books [26, 27]. Also, books on more specialized topics such as social networks [12, 16], biological networks [28, 29] and dynamics on networks [30] are available.

First, we will go through fundamental notation and concepts. This is followed by a discussion of node degrees and their statistical distributions. Data collected for Publication IV will be used here to exemplify the central statistics and their typical behavior in complex networks.

2.1 Basics

A network, or a graph $G(V, E)$, consists of a set of nodes, or vertices, V, that are connected by a set of links, or edges, $E \subset V \times V$ [31]. A convenient way of representing networks is the adjacency matrix, commonly denoted with the letter A. It is a binary matrix, where the element $A_{i,j}$ takes the value of 1 or 0, depending on if the nodes i and j are connected or not. In this Thesis, we are dealing with undirected networks: if there is a link from i to j, this also implies a link from j to i, and we can simply say that there is a link between i and j. This implies that the adjacency matrix is symmetric. In addition, we do not allow self-links, *i.e.* links from a node to itself. In an adjacency matrix of a network without self-links all the diagonal elements have value of 0, *i.e.* $A_{i,i} = 0, \forall i$.

2.1.1 Degree distributions and correlations

The degree k_i of node i is the number of connections between i and other nodes. With the adjacency matrix, the degree is defined as a row (or column) sum $k_i = \sum_j A_{i,j}$. Typical large-scale empirical networks are sparse, which means that most of the adjacency matrix elements are set to 0. This is because the degrees, and thus also the average degree $\langle k \rangle = \sum_{i,j} A_{i,j}$, are under some physical constraints that do not depend on network size. For example, in a social network the number of people one knows is not affected much by the total number of people in the world. On a random network (*i.e.* Erdős-Rényi network [13]), the degree of a node follows the binomial distribution that converges to a Poisson distribution if the average degree is kept constant when the network size grows. In both of these distributions, the degrees are centered around the average degree $\langle k \rangle$ and large deviations from the average are extremely unlikely.

For a long time, the degree distributions of empirical networks were not studied, and thus their forms were not known. However, since ER networks were considered as adequate models of many empirical networks, it was implicitly assumed that the degree distributions should resemble binomial distributions. In 1999, Barabási and Albert [4] published an article challenging this simplistic view: they showed that actually, many networks contain hubs, *i.e.*, nodes with such a high degree that they should be practically impossible to observe in ER networks. More precisely, they found out that networks were scale-free: the degree distributions followed power-laws (*i.e.* $P(k) \sim k^{\alpha}$) instead of being exponentially decreasing

as for ER networks. After this finding, many other networks [22], such as larger samples of the Web [32, 33], the Internet [34], protein interactions [35], and metabolic networks [36, 37] were reported being scale-free. Later on it was found using rigorous statistical testing that some of the networks previously reported as scale-free actually do not have a power-law degree distribution, or the functional form of the distribution is unclear [38, 39]. Also other forms of degree distributions such as power-laws with cut-offs due to some constraints have been found [40]. Nevertheless, most degree distributions of real-world networks are still fat-tailed, *i.e.* the nodes are highly heterogeneous in degree.

The power-law (or fat-tailed) degree distribution is an important feature of a network since it affects the dynamics taking place on top of the network [41], makes the network resilient to random failures but vulnerable to attacks [42], and suggests preferential attachment as a possible model for the evolution of the network [4]. The value of the exponent of the power law affects the qualitative behavior of these features [41], and thus the power-law exponent of the tail is sometimes reported to give an indication how fat the tail is, even in cases where the degree distribution may, in fact, not be a power law.

Figure 2.1 displays the degree distributions of various types of empirical networks (see Publication IV for details of the data). It is remarkable how much the degree distributions of networks obtained from different contexts resemble each other. All of them are undoubtedly fat-tailed. However, the power-law nature of the distributions is questionable already with a visual inspection, since a power-law distribution should show as a straight line in a double-logarithmic plot.

In some networks, such as social networks, there is a tendency for nodes of similar degree to be connected to each others, *i.e.* there is *assortative mixing* [43, 44]. Other networks, such as communication or biological networks, display disassortative behavior, where the hubs are mainly connected to low-degree nodes. The assortitivity of a network affects its vulnerability to attacks, error tolerance and spreading dynamics [43]. The assortativity can be measured by a correlation coefficient considering the degrees of connected vertices, or simply by plotting the average neighbor degree k_{nn} [45] as a function of the node degree. The latter approach is depicted in the inset of the Fig. 2.1, where it is seen that networks associated with information diffusion are organized in highly disassortative way, contrary to the social networks where the popular people are likely

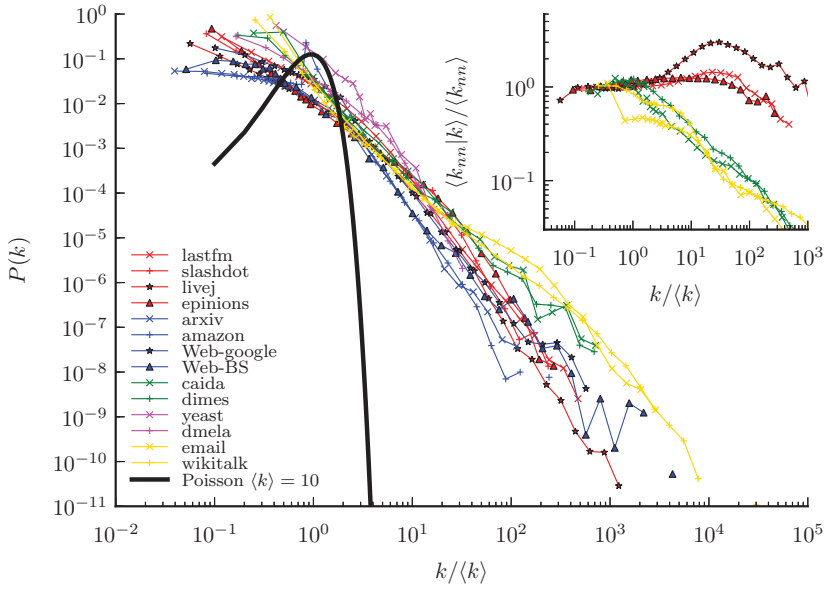


Figure 2.1. Degree distributions for social (red), information (blue), Internet (green), biological (magenta) and communication (yellow) networks. See Publication IV for details on the networks. The black line corresponds to Poisson distribution with average 10, which would results from putting the edges at random to the network. The inset displays the average neighbor degree as a function of the average degree (both normalized with the average values of the networks).

to know each other.

2.1.2 The clustering coefficient

In many empirical networks, the local neighborhoods of the nodes are highly connected, *i.e.* the networks contain much more triangles than would be expected in an uncorrelated random network. The tendency of nodes to form *clusters* is well known in the social network literature [12, 16]: it is clear that two of your contacts are more likely to know each other than two randomly picked persons. This tendency is quantified by the *clustering coefficient* [3] C_i which is defined as the the probability of two neighbors of i being connected. That is, the clustering coefficient counts the number of triangles t_i around the node i, and normalizes that number by the maximum possible number of triangles around a node with the degree k_i. We can write $C_i = t_i/\binom{k_i}{2} = 2t_i/[k_i(k_i-1)]$, or with the adjacency matrix $C_i = \sum_{k,j} A_{i,j}A_{j,k}A_{k,i}/[k_i(k_i-1)]$. $C_i = 1$ if all the neighbors of i are connected and $C_i = 0$ if none of them are connected. In an ER network with link probability p, the expected value for the clustering coefficient $C_i = p$.

The clustering coefficient is typically negatively correlated with the node degree, as is depicted in Fig. 2.2a, where the average clustering coefficient is plotted as a function of the degree for several empirical networks. There have been attempts to explain the clustering spectra observed in data by hierarchical [46] or pseudofractal [47] network structure, which would result to the average C being inversely proportional to the degree: $C \sim k^{-1}$. As is evident from the Fig. 2.2a, at least the set of example networks used in this thesis do not strictly follow this law. Furthermore, a clustering coefficient inversely proportional to degree may be obtained with very simple mechanisms, such as building a network from triangles, and thus it is not necessarily related to hierarchy or fractality.

It is natural that the clustering coefficient decreases with the degree in networks that are sparse and have fat-tailed degree distributions, since the normalization of the clustering coefficient penalizes heavily on large degrees. Let us take as an example the Wikitalk network of Fig. 2.2, where the largest hubs have degree of around $k = 10^4$ and the average clustering coefficient is around $C = 0.2$. For the largest hubs to have clustering around the average clustering coefficient, there should be $Ck(k-1)/2 \approx 10^7$ triangles (or edges) around each node. However, there are only approximately 4.6×10^6 edges in the whole network.

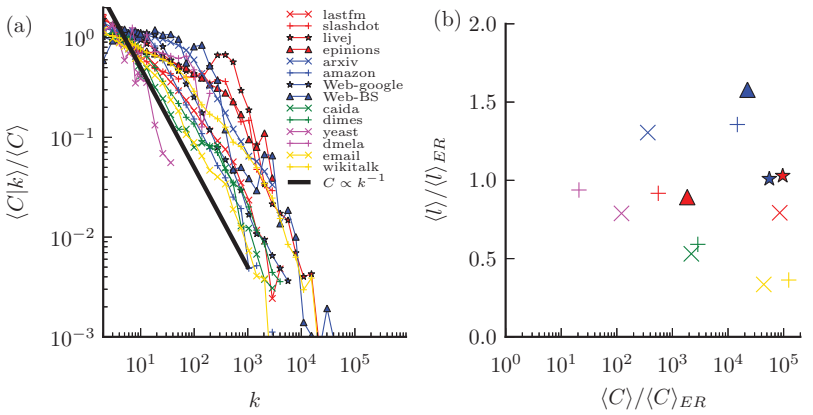


Figure 2.2. *(a)* The clustering spectrum for social (red), information (blue), Internet (green), biological (magenta), and communication (yellow) networks. See Publication IV for details on the networks. The black line corresponds to the case where the clustering is inversely proportional to the degree. The clustering spectrum of an ER random network would be a horizontal line. *(b)* The average clustering versus the average path length for each of the empirical networks. Both of the average values are normalized with the corresponding expected values for the ER random graph, *i.e.* $\langle C\rangle_{ER} = \frac{2L}{N(N-1)}$ and $\langle l\rangle_{ER} = \frac{\log N}{\log 2L/N}$

2.1.3 Paths and distances

A *path* in a network is a sequence of connected vertices, and the number of links on a shortest path from a node to another is the *path length* or *distance* between the two nodes. The longest distance observed in a network is called its *diameter*. Path lengths have a large effect on dynamics on top of the network. Take a social network as an example: If the distances in the network are large, it is hard for any new information, disease or opinion to spread on the network since it has to pass many people in order to reach other parts of the system. In fact, the path lengths set an upper bound for the speed of dynamics on networks.

2.1.4 Small-world networks: high clustering, short paths

Early on, networks were mainly modeled either as regular lattices or random networks, depending on which kind of structure one wanted to emphasize: regular lattices are clustered and random networks have short distances. In 1998, Watts and Strogatz published an article [3] where they introduced a set of empirical networks having both of these qualities. They called such networks small-world networks after the small-world problem in social sciences [48, 49]. The article also demonstrated how lattices can be turned into small worlds by randomly rewiring a small number of links. Later, almost all empirical networks have been shown to be small worlds, which was conjectured by Watts and Strogatz. This is hardly surprising as only a few "random" links are enough to reduce the shortest path lengths of any network to the required level. The small-world article was highly important for complex networks research, since it turned attention away from the simple lattices and entirely random networks, and encouraged focusing on the topology of real-world networks instead of making such simplistic approximations.

In order for a network to be classified as a small world, it must display the short path lengths of ER networks ($\langle l \rangle \propto \log N$), but still have the high clustering coefficient of the regular networks, *i.e.* show randomness on top of regularity. Fig 2.2b shows the average clustering coefficient and average path length values as compared to those of ER random networks for the selected set of empirical networks. The clustering is much higher in all of the networks than would be expected from the random network. The average path lengths are equal or lower than for the random networks with the exception of three of the information networks. The three

information networks – Arxiv, Amazon, and Web-BS networks must have heavy clustering or other structural coherence reducing the number of "random" shortcuts that decrease path lengths.

2.2 Percolation – connectivity of networks

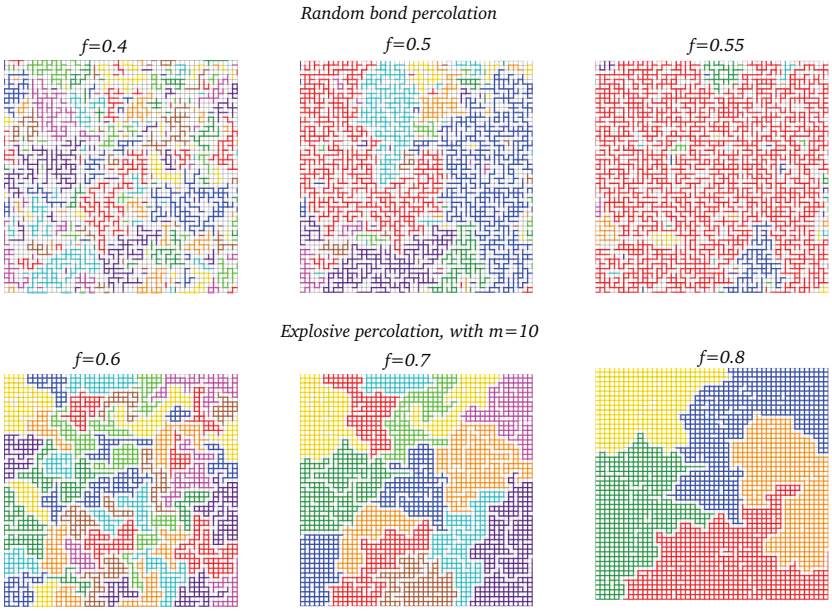


Figure 2.3. Bond percolation on a square lattice. Occupied edges are thick and colored and unoccupied edges are thin and gray. Each percolation cluster has a color assigned to it (colors are repeat if there are many clusters). For the *random bond percolation*, the percolation threshold is $f_c = 0.5$. At the occupation probability $f = 0.4$, all clusters are small, and at the threshold level $f = 0.5$ there are many large clusters spanning almost the whole system as well as a number of small clusters. Just after the threshold at $f = 0.55$, there is a single giant component in the system.

Percolation theory [50] studies the connectivity of networks. Historically, the considered networks have been simple lattices, where some of the nodes (site percolation) or edges (bond percolation) have been deleted (or set unoccupied, in percolation theory jargon). The nodes can then be assigned to connected components (clusters), where there is a path between each pair of nodes. The size of the clusters is an important question in many real situations. For example, oil fields can be modeled as 3D lattices representing the ground where some of the nodes are solid rock, and others have pores filled with oil that can percolate freely between neighboring pores. That is, when drilling a hole to the rock, one can only pump out the oil from all the nodes belonging to the cluster that one hits. If the

system contains large clusters then it becomes economical to drill the oil.

Usually the nodes or edges are considered to be occupied at random independently of each other with some probability f. Consider for example the square lattice network of Figure 2.3, where each edge is set either occupied with probability f or unoccupied with probability $1-f$. If f is small enough, the system consists of a large number of small clusters, and if it is large enough, most of the nodes in the system are connected to a giant cluster. What is interesting is what happens between these two states: the system changes rapidly from the unconnected state to the connected state when it reaches critical percolation probability, f_c, which in the case of bond percolation on a square lattice is $f_c = 0.5$. Also, at f_c the diameter of the clusters explodes and the cluster size distribution becomes a power-law for large systems.

Percolation theory is closely related to random graph theory [13, 14], since from the percolation point of view a random graph is a bond percolation problem on a lattice where all the nodes are connected. Percolation theory is important for empirical networks, where it provides information on how many nodes or links can be deleted before the network falls apart into separated components [51]. This is important for example for disease spreading on top of a social network [52], where the site percolation transition point indicates how many people should be vaccinated before herd immunity is reached. Percolation theory can be used in development of more efficient vaccination strategies than simply giving the shots to random people. In the Internet, it is important to know how many random failures to the routers (site percolation) and connections (bond percolation) between them can be handled before the whole network becomes disconnected. In more general terms, percolation theory explains why in most empirical networks almost all the nodes belong to the same component.

To quantify the percolation process, some terminology and mathematical machinery from statistical physics is employed. In order to calculate anything analytically, it is usually assumed that the system is of infinite size. First, we need to define an order parameter that has a value of zero when the network is in an unordered phase with no giant component, and a value larger than zero when there is a giant component. We define the percolation probability P as the probability of hitting the largest component if we pick one node at random from the system, *i.e.* $P = s_{\max}/N$, where $s_{\max}$ is the number of nodes in the largest component, and N is the

total number of nodes in the network. We also need to quantify the change in P if the occupation probability p is varied a little. For this, we define the *susceptibility* χ as the expected change in the giant component size if a single link connected to it and a random node of some other component is set as occupied. That is, $\chi = \sum_{s \neq s_{\max}} n_s s^2 / \sum_{s \neq s_{\max}} n_s s$, where n_s is the number of clusters of size s. The values of P and χ as a function of f are shown on Fig. 2.4 for bond percolation on square lattice. The change from the disordered phase to the ordered phase is rapid and the phase transition point is clearly visible as a peak of the susceptibility curve.

The study of *critical phenomena* [25] that take place close to the percolation threshold f_c is an essential part of percolation theory. As an example, at the threshold value, f_c, the tail of the cluster size distribution takes a power-law form $P(s) \sim s^{-\tau}$, and close to f_c the relative giant size scales as $P \sim |f - f_c|^{\beta}$. In these scaling laws, the critical exponents τ and β depend on the dimensions of the underlying lattice [21] or on the degree distribution of the network [53–55].

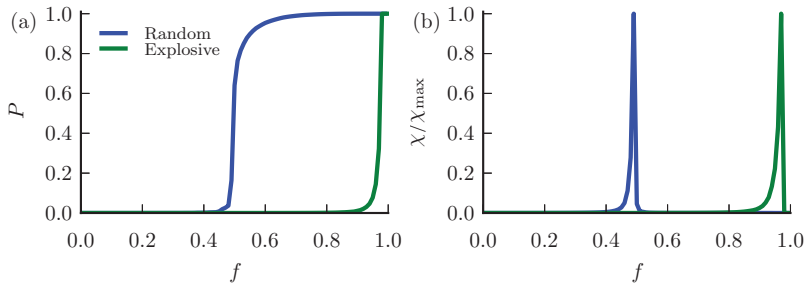


Figure 2.4. Statistics of bond percolation on square lattice. Lattice size is 500x500. Curves for both random bond percolation and explosive percolation are shown. *(a)* The percolation probability (relative giant component size) P as a function of the occupation probability f. *(b)* The susceptibility as function of f. The percolation transition point for the random bond percolation for systems of infinite size is $f_c = 0.5$, but due to finite size effects, the observed susceptibility peak is located at a slightly lower value of f.

2.2.1 Explosive percolation

The concept of explosive percolation was introduced by Achlioptas *et al.* in 2009 [56]. The idea is that instead of studying bond percolation on a lattice or a network with random bonds set as occupied, the occupation is determined by the following irreversible process: the process begins at a point where all the links are unoccupied. Then at each step of the process, m unoccupied links are selected at random. For each of those links $i - j$, the sizes s_i and s_j of the clusters where the endpoint nodes belong to are

determined. Then for each of the links, a score is calculated based on the cluster sizes, and the link with the smallest score is set occupied. Typical rule for calculating the scores is to sum the sizes of distinct clusters $s_i + s_j$ or set the score to zero if both end points of the link are in the same cluster. This step is repeated until all the links are occupied. The process tries to avoid creating large clusters [57], but tries to keep all the clusters roughly equally sized before the percolation transition [58]. This is illustrated in Fig. 2.3 for a square lattice.

What made explosive percolation theoretically appealing was the fact that it seemed to result in a discontinuous (first order) phase transition instead of a continuous (second order) phase transition which is typical to percolation problems. In addition, the explosive percolation transition displayed scaling which seemed to be in contradiction with it being a first order transition. However, it was later argued that the transition is indeed continuous [59, 60], but just very sharp ($P \sim |f - f_c|^\beta$, with $\beta \approx 0.0555$) [59].

The explosive percolation process is studied mostly in artificial networks such as regular lattices [61] and scale-free networks [62, 63], but very little work has been done for empirical networks [64]. In Publication VI we asked what would happen if we applied the explosive percolation process to some empirical network. To our surprise, we found two universality classes, that depend on network topology by observing the behavior of the explosive percolation process. We also found out that the explosive percolation process tends to first accumulate the communities of the network into separate clusters before joining them at the percolation transition.

2.3 Communities

Until now we have discussed either local network characteristics, such as degree or clustering coefficient, or system-level behavior such as the percolation transition. However, empirical networks contain rich structure somewhere between the local and global scale: social networks are organized into families, groups of friends, and even larger units such as institutions or companies. Genetic relationship networks of individuals contain populations separated from each other by physical barriers [67]. Web pages with similar content tend to link to each other [68]. In network terms, all of these are examples of communities: dense subgraphs or clusters found inside networks (see the illustration of Fig. 2.5). These can,

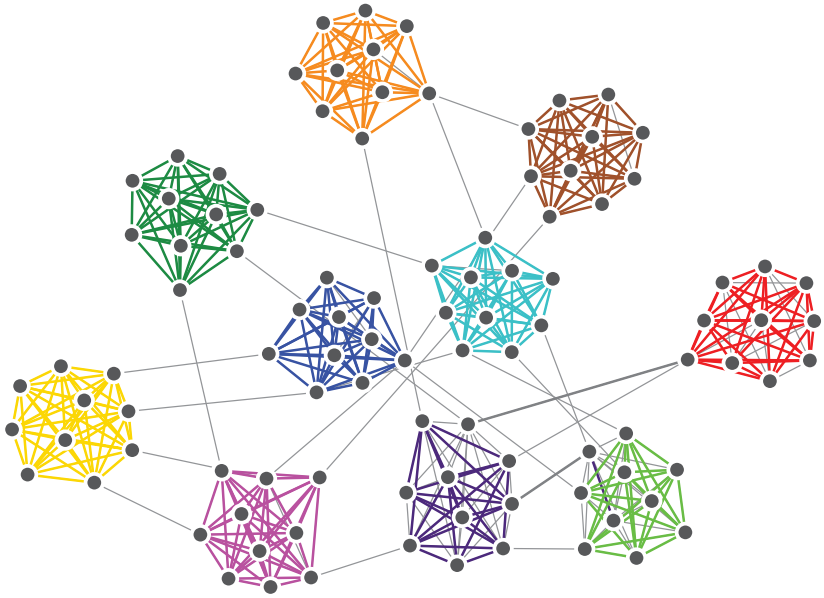


Figure 2.5. A network with communities. A network produced with a simple network generation model with communities [65, 66]. The communities are detected using the explosive percolation method explained in the Publication VI.

in theory, be detected by observing the topology of the network without *a priori* knowledge of the qualities the nodes in the communities share.

The immediate practical problem one faces when trying to find the communities is that there is no exact, universally accepted definition of what is a community. Instead there is a wide variety of mathematical definitions and related computational methods for community detection, all of which have a common goal of finding dense subgraphs in empirical networks [69]. Hence, there is no free lunch when it comes to community detection: Some prior knowledge about the properties of the communities must be available in order to choose the most suitable community detection method.

Some properties such as if the communities can overlap or not might be clear from the context. For example, in most social networks a person can clearly belong to more than one community such as work and family [70, 71]. However, it is not always clear what properties of communities each method emphasizes, since community detection methods are often defined as algorithms instead of defining the communities starting from axioms that tell what the communities should be like. Most community detection methods sound reasonable, but can still find completely different community structures [72]. A similar paradoxical situa-

tion was encountered in the related field of data clustering much earlier than community detection became popular in complex networks. This situation was made clearer by a theorem stating that no such clustering method exists which would satisfy four very reasonable axioms about the clusters [73]. Instead, each clustering method is simply a compromise between the properties a method can have. A similar theorem for communities or clusters in graphs is still to be found, and the emphasis on community detection is on asking which community detection method is better instead of asking at kind of communities a method can find [74, 75], although overlapping communities are already acknowledged as special case in community detection [71, 76–81].

Probably the most well-known method for community detection is modularity optimization [82]. Modularity is a global measure for the quality of a given community structure on a network. It is defined as the difference between the fraction of edges inside the communities (as opposite of edges between the communities) and the expected fraction of such edges in a randomized network. The communities are detected by finding the community structure which gives the highest value for the modularity. As the number of possible community structures grows exponentially with the network size, the optimization is usually done only approximately by some heuristics, such as greedy agglomerative optimization [83, 84], genetic algorithm [85], spectral methods [86, 87], simulated annealing [88], rounding of integer linear programming problems [89], and approximation algorithm [90], where the two latter ones can even give some guarantees for the error [90]. In addition to being best known, the modularity optimization method is also the one which has the most known disadvantages: it requires solving an NP-complete optimization problem [91], there are usually many only slightly suboptimal solutions which are far from each other [92], there is a resolution limit which depends on the network size and sets a lower bound for the size of communities that can be found [93], and the method finds communities in random networks [88, 94]. Luckily, there are plenty of alternative methods for community detection. One of the methods which is gaining popularity and was found to perform well in a recent comparison [76] is the Infomap [95, 96]. It is similar to modularity optimization in a way that it tries to optimize a global quality function for a partition of nodes, with the difference that the quality function is defined as the expected description length of a random walk in the network. One of the advantages of this definition is that it is flexible, such that it

can be generalized to allow overlapping communities [97] and hierarchical community structure [98].

Despite the problems in community detection, the community structure found from empirical networks seems to have universal properties, much like the degree distribution, clustering and path lengths, or show signature characteristics depending on the class of networks, similar to assortativity. The characteristics of communities found in different classes of empirical networks are discussed in Publication IV.

2.4 Network models

Until now we have discussed the structural properties of networks and found out that many of the network characteristics display similar behavior in most networks, even though these come from different sources. However, we have not discussed what could have lead to such "universal" behavior of complex networks. This question is actually composed of two questions: First, are some of these properties only the consequences of some other properties? And second, is there some simple procedure of graph generation that would produce all these properties?

Both of these questions can be answered by network models. The first one is addressed by reference models that are ensembles of networks that contain some selected set of properties of the empirical network but are otherwise completely random. The second question can be approached by simply constructing generative models approximating some aspects of the mechanisms behind the evolution processes of the observed networks. The next two subsections will briefly introduce both types of models, beginning with the reference models.

2.4.1 Reference models

Reference models can answer questions such as "is the property X in my system the direct reason for some other property Y that the system has?" For example, can the degree distribution in a network explain the short path lengths in the same network? Technically, a reference model is defined as an ensemble of networks $\mathcal{G}$ from which networks are sampled with probability $p(G)$. There are two main ways of defining that $\mathcal{G}$ has the set of properties X_i. The first is the "microcanonical" ensemble where all networks are included that display exactly the same values for the prop-

erties X_i observed in the empirical network. In other words, $p(G) = 0$ if $X_i(G) \neq X_i(G_e) \forall i$, where G_e is the empirical network and X_i is the function returning the ith property X_i. Otherwise the ensemble is maximally random, which means that all other probabilities are equal to the inverse of the number of possible networks. The second alternative is the "canonical" ensemble, where one loosens up the conditions in a way that the ensemble displays the properties only on average: $\langle X_i \rangle = X_i(G_e)$, where the averages are defined as $\langle X_i \rangle = \sum_{G \in \mathcal{G}} X_i(G) p(G)$. Here we again want the ensemble to be otherwise maximally random, which can be achieved by selecting the ensemble fulfilling the constraints with maximal entropy [99–101]. The probability distribution for such an ensemble can always be written in exponential form $p(G) \propto e^{\sum_i \lambda_i X_i(G)}$, where the constants λ_i depend only on the values of the properties on the empirical network $X_i(G_e)$.

The simplest reference model is the ER random graph [13, 102] that we have already discussed and used several times in this Thesis. The microcanonical version of the ER model keeps the total number of links L constant in all networks of the ensemble, *i.e.* it simply shuffles the links. In the canonical version of the ER model[1], each pair of nodes is linked with a fixed probability $p = L/\binom{N}{2}$, where N is the number of nodes in the network. Both versions of the model give roughly equal results, but the canonical version is usually used for any analytical results and the microcanonical for simulating sample networks, *e.g.* when shuffling empirical data. As discussed earlier, the ER model cannot explain the heterogeneous degree distribution of empirical networks. A random reference model where the degree distribution is retained is called the configuration model [53, 105–108]. The microcanonical version of the configuration model is more popular, but the canonical version is also sometimes applied [100, 101].

In social network analysis, the canonical versions of the reference models have been used for a long time under the name of exponential random graphs, or p^* models [109–113]. Typically, exponential random graphs in social network literature are used as sophisticated reference models that can include properties (having fixed ensemble average value) related to the number of triangles, degree, and even external attributes such as age,

[1]Erdős and Rényi didn't consider the canonical version of the random graph model in their seminal article considering random graphs [13]. However, both versions of the model are commonly named after Erdős and Rényi [14] even though there were several other authors with similar ideas already before them [103, 104].

sex or race. The downside of using such a complicated model is that the analytical handling of these models is usually impossible and even sampling from the ensemble and fitting the model to data can be hard [113–115].

2.5 Generative models

Generative network models are designed to answer the question of what kinds of mechanisms could have produced the observed network structure. Of course, the fact that one network creation procedure can lead to the observed network doesn't remove the possibility that there could still be some other mechanism that can do the same. To address this concern, network generation mechanisms are usually designed to be as simple as possible, since a simple explanation for the network evolution is more believable than a complex one. Also, in most systems there are several known mechanisms molding the networks, but their relative importances to the overall structure of the networks are not known. The typical modelling approach is then to take one or two of these processes, simplify them, and try out building networks according to them.

The fat-tailed degree distribution observed in most of the empirical networks cannot be explained by the ER network model. Instead, there must be some sort of mechanism behind the network creation process that is so simple that it is plausible for all different systems. The preferential attachment process, where networks grow by new nodes linking preferentially to high-degree nodes, fits these criteria perfectly. This mechanism of network evolution was first described by Price in 1976 [116] in the context of citation networks, but was later made popular by Barabási and Albert [4] when explaining the ubiquity of the scale-free networks.

One of the most modeled types of systems in complex networks literature are social networks, which in addition to having the fat-tailed degree distribution typically display characteristics such as assortativity and strong community structure. In the Publication III, we review and compare several social network models and assess their ability to produce realistic social network topology. We divide the generative network models to network evolution models and nodal attribute models, and also include exponential random graphs for comparison.

3. Weighted networks

The previous Chapter dealt exclusively with networks where each pair of nodes is either connect or not, and all other information about the systems is discarded. As we have seen, this approach has been extremely successful in describing a variety of systems, and has lead to the discovery of many almost universal properties of empirical networks, some of which are emergent and can be explained by simple network evolution mechanisms. However, for many classes of networks, the simple topology is not enough, but the interaction strengths, or edge *weights*, display high heterogeneity, making them an essential part of the system. As an example, for social networks it might not be enough to take into account who knows who, but how well people know each other matters [6], and in transportation networks the capacities of the connections are an important factor [5, 117, 118].

The way of defining interaction strengths, or weights, varies a lot between different systems, as is evident from Table 3.1 that lists some of the weighted data sets appearing in networks literature. Similarly to unweighted networks, where various types of interactions between the elements are reduced to links between the nodes of a graph, we can map all the different types of interaction strengths to positive real numbers: an unweighted graph $G(V, E)$ may be generalized to account for the weights by defining a weighted graph $G(V, E, w)$, where V and E are the sets of vertices and edges as before, and w is a weight function from the edge set to (positive) real numbers ($w : E \to \mathcal{R}_+$). Again, similarly to unweighted networks, where the network structure is represented with an adjacency matrix, weighted networks can be represented by a weight matrix W, where $W_{i,j}$ is set to zero if there is no link between the nodes i and j, and otherwise it is set to the weight of the edge between the two nodes $W_{i,j} = w(i, j)$.

Network name	Node	Weight	Citations
Mobile phone	Person	# of calls	[10, 119, 120]
International trade	Country	Trade volume	[121, 122]
Air traffic	Airport	# of seats	[5, 40, 118]
Scientific collab.	Scientist	# of coauth. articles	[5, 122–125]
Brain connectivity	Brain area	# of fibers tracts	[126–128]
Brain connectivity	Brain area	prob. of fiber tract	[126]
Stock market	Stock	Stock price cor.	[129]
Genetic network	Gene	Genetic distance	[130, 131]
Emotion words	Word	Similarity from 0-5	[132]
Clique overlap	Clique	# of shared nodes	[79]
Metabolic network	Chemical	Flux of reaction	[133]
Food web	Species	Flow of biomass	[134]

Table 3.1. A sample of weighted network data sets. For each network, the type of node and edge weight is listed.

For some systems, representing the interaction strength as a real number is straightforward, *e.g.* when there are multiple connections between pairs of nodes [65], when links have well defined probabilities of existing [126, 135], or when there is a flow or a flux between pairs of nodes [133, 134]. However, in some other systems the interaction strengths are not one-dimensional, but the link weights need to be defined by starting with some higher-dimensional data, and calculating some summary statistics reflecting the strengths of the interactions. If the links have a temporal dimension, such as in mobile phone call networks [119], it is usually averaged out, *e.g.* by considering the average rate of calls as the link weight (there are also alternative statistics [136]). Bipartite networks [31, 124, 137, 138] can be projected into weighted networks by calculating the (normalized) number of shared neighbors of nodes, such as in scientific collaboration networks, where the weights correspond to the number of coauthored articles [5, 122–125]. Note that many of the collaboration networks also have a time dimension, *e.g.* the release dates of the movies and the articles. Finally, there are systems where the nodes, instead of the interactions, are high dimensional feature vectors. Such networks include stock market networks, where the nodes represent the time series of stock prices, and the weights are the correlations between them [129], and genetic networks, where each node corresponds to a geno-

type of an individual and the weights are genetic distances between the genotypes [130, 131]. The distances are usually further transformed to similarities, since a high distance corresponds to a low tie strength. This way of defining weighted networks results in full networks, where all the possible edges are present.

The fact that all weighted networks can be represented with the same abstraction implies a possibility to repeat the success story of the unweighted networks: to develop weighted network metrics [5] and apply them to all possible weighted systems in order to observe some ubiquitous behavior or find classes of systems with similar properties. However, it turns out that weighted networks are not that simple. Consider for example two networks, one where the weights are defined as probabilities for the edges to exist, and another, where they represent distances between the two nodes. The relevant questions one can ask about these weighted graphs differ greatly: in the probability graph, one might *e.g.* ask how likely it is for some clique to emerge in the network, which could then be calculated as a product of the probabilities of the edges of the clique. In the distance graph, this calculation doesn't make any sense: one simply cannot use the same set of tools for all weighted networks since the interpretations of the weights are different.

The question that arises naturally is whether the weights can be somehow categorized, such that these categories could be used to decide what tools are appropriate for analyzing different networks. Similar questions have arisen in the field of statistics, and in the 40's there was an attempt by Stevens [139] to build a hierarchy of measurement scales, and to classify the statistical procedures according to what type of data is appropriate for what scale. As an example, it is typical to assign numbers to data that is actually on an ordinal scale, *e.g.* at the scale of good (1), neutral (0) and bad (-1). According to Stevens it would be forbidden to calculate any statistics of the data, such as the average, that would be affected if some other similarly ordered numerical values were assigned to the data classes. Stevens' approach has attracted a lot of critique [140, 141], because defining the scales for empirical data can be problematic, and following such categorization too strictly can unnecessarily limit the data analysis. The type of analysis that can be performed on the data is not determined by the type of variables, but by the type of questions the researcher has regarding the data. The same is true for weighted networks. Let us take as an example a mobile phone call network, where the num-

bers of calls are used as the edge weights. If you are a mobile phone operator you will probably consider the number of calls as absolute scale and ask questions such as "how many more calls per month users of type X make than users of type Y?". However, if you are a sociologist interested in the underlying social network, and consider the number of calls as a proxy for the tie strengths between the people, then you probably consider the number of calls in an ordinal scale, where more calls means a stronger tie, but you don't ask questions such as "By how many calls per month is person A better friends with B than with C?".

This Chapter introduces two cases with two different approaches to tackle the problems in analysis of weighted networks discussed above. In the first case, some possible ways of generalizing the clustering coefficient for weighted networks, and the problems related to them, are reviewed. It turns out that the heterogeneity in the ways of defining weights for different systems has lead to variation in the interpretation of what weighted clustering means, and ultimately to different generalizations for the weighted clustering coefficient. In the second case, community detection in weighted networks by applying thresholding and clique percolation is discussed. Thresholding-based methods are non-parametric in the sense that they only consider the order of the weights, allowing one to use them for a wide variety of systems.

3.1 The weighted clustering coefficient

One way of approaching weighted network analysis is to take some unweighted network characteristics and generalize them so that they take weights into account. Perhaps the most successful of such generalizations is the node *strength*, which is an extension of the degree [5]. The strength s_i is simply the sum of the weights of the node, $s_i = \sum_j W_{i,j}$, similarly to the degree which is a column (or a row) sum of the adjacency matrix. In most networks, the sum of the weights is a natural measure of the importance of the node: in some networks the sum represents the total flow, count or rate related to the focal node. In networks weighted with the edge probabilities, it equals the expected number of edges the node has.

There have been several attempts to generalize the clustering coefficient, probably because it is one of the most well known and widely used network characteristics in the complex networks literature. In Publication I, we review four of these generalization attempts. Perhaps sur-

prisingly, the different generalizations of the clustering coefficient behave quite differently from each other, even though all of them have the same goal of measuring local clustering around a node. On the other hand, this is not surprising at all since the generalizations have been designed with different types of systems in mind, and thus the questions behind them are different. Barrat *et al.* [5] were interested in airport and collaboration networks, and for these the relevant question was how much of the relative weight of the links of a node is associated with triangles, *i.e.* the clusters. For Grindrod *et al.* [142] and Ahnert *et al.* [135], the starting point was networks where edge weights are defined as the probabilities that the edges exist. Here, a natural question to ask is what the expected value of the unweighted clustering coefficient is. Both defined the weighted clustering coefficient as an approximation of this expected value. The same formula for the weighted clustering coefficient can be reached by replacing the adjacency matrix with the weight matrix in the most commonly used formula for the unweighted clustering coefficient [143, 144]. Onnela *et al.* [145] defined the weighted clustering coefficient as a product of the average of the intensities (*i.e.* geometric means of the weights) of the triangles around the node and the unweighted clustering coefficient. This compromise between clustering of weights and topological clustering was seen to work well for stock correlation networks, and produced reasonable results even when the weighted network was fully connected. Thus, none of the generalizations is better than the others – the questions they answer are simply different.

In Publication I, it was noted that all of the weighted clustering coefficients are compromises between the topological clustering around the focal node and the organization of weights around that node. This makes all of the coefficients highly degenerate, and thus hard to interpret when empirical data is used. Interpreting results for the unweighted clustering coefficient already requires taking into account information on the node degrees. Adding a new dimension, the weights, to the equation is not going to make interpretation any easier. That is, if the value of the weighted clustering coefficient of a node is low, the reason can be any combination of the following: (1) there are not many triangles around the focal node, (2) the degree of the node is high, which in many networks with otherwise high clustering coefficient values limits possible available values for the clustering (see discussion in the previous Chapter), or (3) there is not enough weight associated with the triangles around the node.

The problems with the degeneracy of the weighted clustering coefficients can be circumvented by asking more specific questions about how the topological clustering and the weights are related. One approach is to calculate intensity or coherence distributions of the triangles [145]. In Publication VII, we wanted to find imbalanced triplets from a network of emotion concepts, or more specifically triplets where two edges have high weights and one has a low weight. We could have used coherence values to find such structures, but triplets with one strong connection and two weak ones also have low coherence. We could have then used intensity values to filter out these unwanted triplets. Since our original question was quite simple, we instead constructed a new measure for answering that specific question.

3.2 Thresholding and clique percolation

A popular technique for studying weighted networks is to use the weights to filter out a part of the edges, after which all usual weighted and unweighted network methods can be applied. The most popular and simplest filtering method is *thresholding* the network by discarding all edges with weight below (or above) a certain *threshold* level. This technique can be applied with a single threshold level, or by sweeping through all threshold levels and plotting appropriate metrics as a function of the threshold. Thresholding-based methods can be, in theory, applied to all networks where the weights can be (partially) ordered. Other filtering methods, such as finding minimum/maximum spanning trees and backbone extraction methods [146], are not considered in this Chapter.

Theoretically, thresholding is close to bond percolation, with the difference that the edges are not set occupied and unoccupied at random, but according to their weight. This correspondence is useful when applying thresholding-based techniques to empirical system, since results for percolation of random networks can be directly used as a reference model for the thresholding [129]. Also, the quantities of interest, such as the giant component size and the susceptibility, can be directly adopted to monitor the progress of the thresholding sweeps. This approach has been used for example to determine a suitable threshold value for a genetic network of populations [130], and to prove Granovetter's theorem [6] stating that the society is held together by weak social links [119, 120].

3.2.1 Algorithmic perspective

The algorithmic cost of a threshold sweep, where the network characteristics are calculated at each threshold level, might sound high at first. However, the sweeps can usually be implemented in an algorithmically efficient way, such that some of the key statistics from all threshold levels can be extracted swiftly even for extremely large networks. The idea is to begin with an empty network, and start adding edges one by one in the order induced by the weights. After each addition, one simply updates the changes in the statics. If the statistics are local in nature, *e.g.* the clustering coefficient in a sparse network, these changes can usually be calculated in constant time. That is, the total computational time used for the addition of edges is $O(L)$, where L is the number of links in the network. The total time for the whole procedure is thus dominated by the sorting time $O(L \log L)$.

The same algorithmic approach is also widely used in percolation analysis, where instead of ordering the edges by weight, they are added in a random order. This procedure produces statistically correct uniform samples from the set of networks with the given number of added edges. However, the samples are not independent, which is usually not a problem since the procedure can be repeated to check for variations in sample networks having the same number of edges. In fact, Fig. 2.4 of the previous Chapter displaying the giant component size and susceptibility at different levels of occupation probability was produced in this way.

In order to calculate the giant component size, one needs to know to which component each node belongs to. Updating this information after addition of each edge requires more than a local inspection of the network. However, this can be achieved faster than $O(\log L)$ by using the disjoint sets forests (DSF) data structure [147]. Further, there is no need to store the network structure in the memory, but only the DSF, which takes only $O(N)$ of memory, where N is the number of nodes. Thus, the whole thresholding procedure can be completed in $O(L \log L + N)$ time.

3.2.2 Weighted clique percolation

The *clique percolation method* (CPM) is a community detection method introduced by Palla *et al.* [70] in the 2005. The main motivation behind the method was to detect overlapping communities, which were overlooked by other community detection methods at the time. The CPM is based on

the observation that communities usually contain cliques, *i.e.* subgraphs where each node is connected to every other node, and that the cliques are closely related to each other. To be more precise, community detection with the CPM starts by transforming the network into a k-clique graph G^*, where the cliques of size k form the nodes, and two cliques are connected if they share $k-1$ nodes on the original network. For example, if $k=3$ the triangles are the nodes and two of them are connected if they share a link. Then, the communities are defined as the union of all nodes in all cliques of each component of G^*.

In addition to the ability of the CPM to find overlapping communities, its main benefits over the older methods are that it is deterministic (it always finds the same communities), and it is easy to understand what kinds of communities it finds, unlike for methods that are, for example, based on solving some optimization problem. The CPM is also a local method in a sense that the community finding algorithm only considers local information and the communities it finds are not affected by changes in distant parts of the network. However, if the clique size k is not selected a priori to detecting the communities, the local nature of the method is partially lost, since the global network density affects the choice of k if the percolation procedure of Ref. [70] is used.

Selecting a proper clique size k is in practice the main problem with the CPM. If k is too large almost no cliques can be found from the network, and if it is too small, there might be a giant community that spans the whole network. If the network is sparse, k should be small, usually around 3 to 6 [70], and because the clique size needs to be an integer, there might not be any good value of k for a given unweighted network. Also, the optimal value of k might be different for different parts of the network, *i.e.* some parts of the network can be more dense than the others. And finally, even if there is a good value of k to be found, it is usually found near the percolation threshold point of the clique network [148]. From percolation theory we know that the clusters near the critical point can be tree-like and have a high diameter, both of which are not desirable properties for communities. The properties of clique graphs formed when starting from ER random networks are already well understood from the percolation point of view [148–150], but the structural properties of empirical clique graphs have not gained much attention [79].

If the empirical network is weighted, it can be thresholded before finding the communities using the CPM. This allows one to select such a small

clique size that it would results in too large communities without thresholding, and then adjust the number of links such that the community structure becomes as rich as possible [70]. As illustrated by Fig 3.1, there is a tradeoff between the weight threshold and the clique size: selecting a large weight threshold (and, thus, a small clique size) results in communities that are based on the strong links of the network, but are not necessarily topologically coherent. If a small threshold level is selected, it is possible to use larger cliques, but communities can be formed between weakly connected nodes. However, if large weights are associated with the cluster structure, as in social networks [6, 119], the network should be robust against the choice of the threshold. Also, a weighted clique percolation method that is based on intensity values of cliques can be used [151].

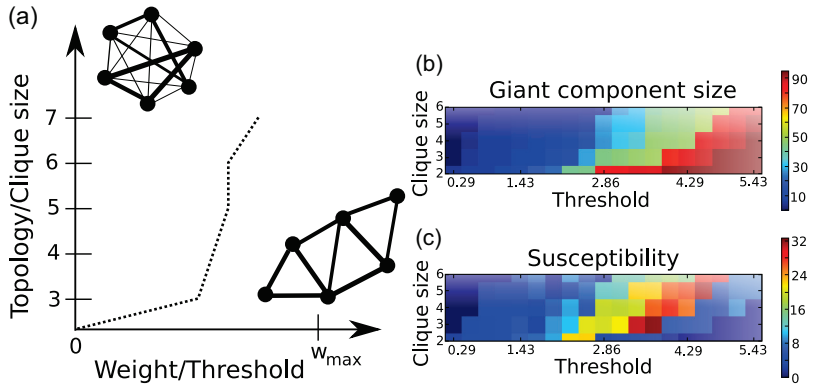


Figure 3.1. Thresholding by topology and weight in clique percolation. (a) A schematic view of the weight-topology tradeoff in clique percolation. Topologically coherent communities can be found by allowing weak links in the communities but increasing the clique size. Communities that have strong weights but are not necessarily topologically coherent are found by increasing the weight threshold while still keeping the clique size small. The dotted line represents the "optimal" threshold value. (b-c) The giant component size and susceptibility (measured in number of nodes in the original network) as a function of the clique size and weight thresholds. The optimal threshold should be so small that there is no giant component, but large enough for cliques to exist: *i.e.* the susceptibility value should be large. The figure is adopted from Ref. [67].

Thresholding does not only solve the problem of finding a value for the clique size k such that the community structure is as rich as possible. It also solves the problem of heterogeneity in the network density that makes different values of k optimal for different parts of the network. In the Publication II we introduce the sequential clique percolation (SCP) method, which is used to build dendrograms displaying the community structure at each weight threshold level, and the way the communities merge when the threshold is decreased. Thus, it doesn't only produce the

community structure at multiple resolutions, but also allows one to find the hierarchical structure of the network.

The power of the CPM and the SCP method is perhaps best illustrated by giving an example. In Publication VII, we investigated a weighted network formed between Finnish emotion concepts using similarities of the concept pairs as weights. Running the SCP method with the clique size set to three for this network produces the dendrogram presented in Fig 3.2. The hierarchical community structure in this network is an essential part of the system, and the benefits of building the whole dendrogram instead of choosing a single threshold are clear: First, there is no clear optimal threshold since selecting one would always leave some interesting structure invisible. Second, the dendrogram reveals the relationships between the communities, such as the positive communities merging together before they merge with the negative ones. An important feature of CPM is the fact that the communities can overlap. Take the three clusters where the concept "happiness" belongs to as an example: depending on the context, the word can have different meanings, which is evident from the overlapping community structure found with the CPM, but would have caused some trouble to any other community detection algorithm not allowing for overlapping communities.

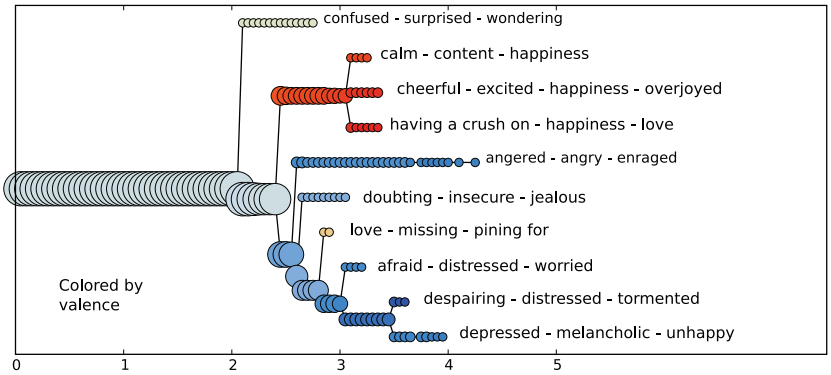


Figure 3.2. SCP applied to the emotion concept network. The circles represent the 3-clique clusters, with the area of each circle proportional to the number of nodes in the cluster. The x-axis denotes the weight threshold level: decreasing the threshold makes the clusters merge such that finally there is only a single giant cluster containing all the nodes. Colors denote the average valence level (positive to negative) of the emotion concepts. See Publication VII for more details.

Algorithmic perspective

A naive solution to the problem of finding clique communities in a network is to explicitly build the clique overlap network G^*, and simply find

the connected components. For a given $k \ll N$, the worst case for this algorithm is the full graph for which the algorithm runs in polynomial time and memory $O(N^{2k})$. However, in real world cases it is often better to build a network where the nodes are the maximal cliques of the graph [70, 152] (*i.e.* cliques which are not completely covered by some larger clique), and link two of them if they share nodes. Considering this network as a weighted network, with weights defined as the number of nodes shared by the maximal clique pairs, all k-clique community structures can be reached by simply thresholding the network (and by considering the isolated maximal cliques separately). The problem with this approach is that finding the maximal cliques is an NP-complete problem [153]. However, there are algorithms that are usually fast for sparse empirical networks [154–156].

The algorithmic approach taken by the SCP method is closer to the naive solution than to the maximal-clique solution. The SCP method improves the naive solution by turning the problem of finding the connected components in the clique graph at each threshold level of the original graph into an edge percolation problem in a modified clique graph. This way, the edge percolation algorithm described in the previous section can be used both to find the new cliques at each threshold level and to update the component structure of the clique graph. In the modified clique graph the $k-1$-cliques are the nodes that are connected if they are a part of the same k-clique, making the k-cliques in effect hyperedges [31] of the graph. The DSF uses less memory in the worst case for the modified clique graph than for the naive clique graph, as the number of $k-1$ cliques is less than the number of k-cliques ($O(N^{k-1})$ versus $O(N^k)$). This is also true for many realistic cases. For example, there are usually more edges, *i.e* cliques of size 2, than nodes, *i.e.* cliques of size 1, in empirical networks. Effectively, the SCP algorithm does thresholding sweeps on the horizontal axis of Fig. 3.1, where the maximal clique algorithm finds all the community structures on the vertical axis. A full description of the SCP algorithm can be found in Publication II.

4. Temporal networks

In the previous Chapter it was argued that due to the heterogeneous nature of interaction strengths in many complex systems, such systems should be represented with weighted networks instead of binary ones. However, if you take a look back at Table 3.1, you will notice that actually, most of those networks are not just weighted – they are also time-dependent. As an example, in the mobile phone call network, the links are active only at the times of the calls, and in the scientific collaboration networks, the number of links is constantly growing as more and more articles are published. Also, brain activation networks [157–159], protein interaction networks [160–163], gene regulation networks [164–166], and stock price correlation networks [129, 167–169] are inherently dynamic in nature.

As with weighted networks, we would like to find a suitable abstraction level at which all different time-dependent networks could be represented. Similarly to weighted graphs, we can extend the static graph representation $G(V, E)$ to temporal graphs $G(V, E, \rho)$ with a *presence function* $\rho : E \times \mathcal{T} \to \{0, 1\}$, where $\mathcal{T}$ is the temporal domain of the system, and the presence function gets value $\rho = 1$ if the edge $e \in E$ is present at time $t \in \mathcal{T}$ and $\rho = 0$ otherwise [170]. Temporal graphs can also be represented with sequences of static graphs or with time-varying adjacency matrices [171]. For some networks it might be necessary to include a *latency function* for the edges [9, 170, 172]. For example, in an airline network a link between two airports is present only at the departure times of the flights with the latency equal to the flight time. In many systems, it is more natural to think of the temporal network as consisting of events [8, 10, 11, 173]. Graphs with presence and latency functions can be equivalently represented with event sequences if we define the events as tuples $e = (i, j, t, d, \delta)$, where i and j are the two nodes participating in

the event and t is the time, d is the duration and δ is the latency of the event.

For the purpose of this Thesis, we will concentrate on temporal networks represented by event lists $\mathcal{E}$ that are sets of events $e = (i, j, t)$ between two nodes $i, j \in V$ at time t. We consider the events to be undirected, such that the order of i and j in the events is not meaningful. Further, we will leave out the duration and the latency. This will restrict our scope to temporal networks where all events have (almost) zero duration and latency, like in email networks, or the duration and the latency are negligible compared to the other time scales, like for mobile phone calls, where the durations of the calls are typically much shorter than the times between calls. Leaving out these extra parameters allows us to concentrate on the timings of events, which are at the core of temporal networks. More complicated temporal networks should be considered only after we understand how the timing of events affects temporal networks.

In this Chapter, the focus is on temporal social networks and dynamics on top of them. If the timings of events in a temporal network come from a uniformly random distribution, the whole temporal networks framework would be unnecessary since weighted, or even unweighted, networks could be used to model such systems. This Chapter begins by reviewing results showing that human behavior in temporal communication networks is highly heterogeneous, instead of uniformly random. We will then discuss the implications of this heterogeneity on spreading processes taking place on top of these social networks. Again, the interested reader is pointed towards a recent review [174] for a more complete picture of temporal networks.

4.1 Heterogeneous activity patterns

The Poisson process is a continuous time model that creates sequences of events such that these take place independently of each other. It is the standard temporal model that has been successfully used to describe various systems from radioactive decay to customers arriving into a queue. However, it is becoming increasingly clear that for many complex systems, from natural ones to ones related to human behavior, the Poisson process doesn't offer a good description, but their dynamics are more heterogeneous instead [175]. This is also true for social interactions and communication in social networks [176]. In this Section, we will go through ways of

measuring and quantifying heterogeneity and bursts in event sequences using a mobile phone call network as an example.

The simplest way of characterizing heterogeneity in any signal consisting of time-stamped events is to calculate the *interevent times* of consecutive events. For temporal networks, one can for example calculate the interevent time distributions of activity patterns of nodes or the activation sequences of links. The interevent time distributions are dominated by the average rate of events, and as we are mainly interested in the heterogeneity of the interevent times, the distributions are further normalized with the average interevent time. Fig. 4.1 displays the normalized interevent time distributions of links in a mobile phone call network [10]. The distributions have been computed conditional to the edge weights (numbers of calls) in order to detect any behavior that would depend on the weights. All the interevent time distributions come from the same fat-tailed signature distribution, independently of the number of calls taking place on the edge. A similar scaling of the interevent times has been also found for the node activations in a mobile phone call network [176].

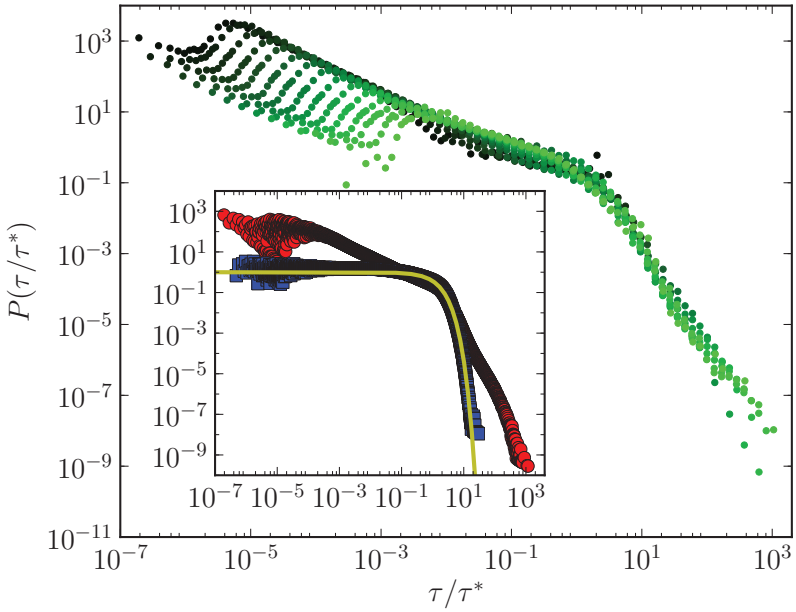


Figure 4.1. Interevent time distributions for calling patterns. Distributions of normalized interevent times of call sequences on top of the edges, conditional to the number of calls on the edge (binned for better statistics for edges with high number of calls). The higher the number of calls, the darker the color of the points in the distribution plot. Inset: the normalized interevent times without conditioning for the original data (in red), for data with only the daily patterns and no bursts (in blue), and for data without daily patterns and no bursts (yellow line). Figure adopted from Publication V.

Even for a Poisson process, there is some variation in the interevent times. Further, the heterogeneity in the communication patterns might be due to the daily pattern alone, *i.e.* because there is more activity during day-time than during the night. To test if the observed heterogeneity in the interevent times is only due to these trivial reasons, we need to build reference models. Similarly to topological reference models, we want the event sequences to be maximally random with the condition that they retain some properties, such as the daily pattern or network topology. Such models are discussed at length in Publication VIII. The inset of Fig. 4.1 displays the normalized interevent time distributions for the original data and for two reference models: one where only the weights and topology of the underlying network are retained, and another where the daily pattern is included in addition. It is clear that the original data is more heterogeneous than the reference models: The original interevent time distribution has a fatter tail, and shows a larger number of short interevent times than the interevent time distributions of the reference models.

It is not always necessary to plot the whole interevent time distribution to quantify the heterogeneity in the time sequences. Goh and Barabási [177] measured heterogeneity by defining a quantity they called *burstiness*, which is the normalized coefficient of variation of an interevent time distribution. The normalization is done in a way that the burstiness will be zero, if the variance of the interevent time distribution is equal to that of a Poisson process with the same average rate of events as in the observed sequence. Further, the normalization limits the burstiness values between -1 and +1, such that higher values correspond to more bursts. Burstiness values higher than zero were reported both for natural time sequences, such as earthquakes and weather patterns, and for temporal social networks [177].

The correlations in event sequences of temporal networks are not limited to the heterogeneity in interevent times. In social contact networks, for example, an event between two nodes A and B might *trigger* an event between B and C [178]. These networks also display larger, mesoscale structures such as *dynamic motifs* [179].

4.2 Spreading and path lengths

Until now, we have not discussed much about dynamical processes taking place on top of networks. However, dynamics such as disease or ru-

mor spreading [24, 30, 41, 180, 181], opinion formation [182–185] and cascading failure [186, 187] are central topics in complex network literature. Some results on spreading models were mentioned when the percolation theory was reviewed (see Section 2.2), since many of these processes are heavily related to paths and percolation. In this Section, spreading models are finally discussed in more detail. At first, we will go through compartmental models [188] with fully mixed populations which form the theoretical basis on top of which more complex spreading models can be constructed. After that, static networks as spreading lattices are shortly reviewed, and finally spreading on temporal networks is introduced.

4.2.1 Compartmental models

Much of the theory of spreading originates from models of epidemic progress in epidemiology. Thus, much of the terminology is referring to epidemics of contagious disease, although one might be discussing, for example, the diffusion of information on social networks. *Compartmental models* are one set of theoretical tools that have been borrowed from mathematical epidemiology. In these models, each individual is in a single state, or compartment, and can move between the compartments according to the rules set by the spreading model. The simplest of such models is the SI model, where each individual is either susceptible (S) or infected (I). Initially, most of the nodes are set to the susceptible state, and susceptible nodes become infected if they come into contact with infected nodes.

In the SI model, all nodes finally become infected, since there is no way the nodes could recover and move back from the infected state to the susceptible state. This is not very realistic, and there are two ways to correct this: First, recovery could take place so that the nodes again become susceptible after some time. This model is usually denoted the SIS model. However, for disease spreading, the patients recovering from an infection usually develop an immunity towards the disease, and thus cannot be infected again. Another alternative is that the patients die, and cannot infect other nodes any further. These situations can be modeled with the SIR model, where nodes in state I move to state R with some rate. Depending on the level of optimism of the researcher, the state R is called either “removed” or “recovered”. There are also other compartmental models which are not considered here. One could, for example, include an incubation period for the disease.

The spreading models, described above are rather simple. The tricky

part is to define when two nodes are in contact such that the infection (or information) can be transmitted to a susceptible node. In traditional epidemiology it is usually assumed that there is no underlying structure in the contacts of the nodes. Instead, each pair of nodes is equally likely to be in contact at each point of time. This assumption is then combined with the assumption that the population is infinitely large, allowing the use of a set of differential equations for solving how the dynamics of the epidemics proceed. For the SI model, the solution to the equation is that the number of infected people grows exponentially at the beginning, and the growth is slowed down only when a significant proportion of the population is infected. Finally, the whole system is in the infected state. For the SIR model, the behavior of the system depends on the *basic reproduction number*. If a person can infect more than one other person on average before moving to the removed state, the epidemic progresses roughly similarly to the SI model, whereas for a ratio smaller than one, the infection quickly dies out. The concept of the basic reproduction number is useful for example in deciding how many people must be vaccinated against the disease in order to lower the basic reproduction number below one, *i.e.* to achieve *herd immunity*. In this case, even unvaccinated people should be safe.

4.2.2 Spreading on networks

Complex networks offer a more realistic way to model the contact sequences than simply considering the whole population as fully mixed. The nodes are considered to be in contact only with those nodes they are connected within the network. The contacts are usually considered to follow a Poisson process with a constant rate over time. If the underlying network has weights, they can be used to determine heterogeneous contact rates for the edges.

The final outcomes of the SI and SIR processes can be mapped into percolation problems in networks [52, 189]. If an SI process is started from a single initially infected node, the infection always spreads to all nodes of the component that the initial node belongs to. For the SIR model, it turns out that a probability can be calculated for each edge, by which the edge transmits the infection if one of the two nodes becomes infected at some point in time. Thus, the size of the infected population is equivalent to the cluster size distribution in an edge percolation problem [190, 191]. The transmission probability of an edge depends on the rate at which in-

fected nodes are removed, on the infection probability, and on the contact rate which can be determined from the edge weight.

The inclusion of networks to generate the contact sequences doesn't only make the spreading model more accurate, but it also moves the focus away from the infection and recovery rates and from the basic reproduction number to the topology of the networks. If the network topology is scale-free, the presence of high-degree nodes, or hubs, causes the epidemic threshold to disappear, such that the entire network always becomes infected [41, 192, 193]. Thus, the hubs should be targets for vaccination, instead of distributing vaccinations to a random subset of people as suggested by the fully mixed population model [194, 195]. Also the distribution of weights has a role in spreading phenomena. For example, in social networks with Granovetter-type structure, the weak links work as bridges between well-connected parts of the network and inhibit spreading [119].

4.2.3 Spreading on temporal networks

Instead of modeling the contact sequences with the fully mixed population model, or as Poisson processes on top of complex networks, we can directly take the contact sequences into account, *i.e.* use temporal networks as spreading lattices. The spreading of diseases, that require some physical contact has been studied with dynamic networks of face-to-face interactions [196], sexual contacts [197–200], hospitals [201], and airline transportation [9]. Information spreading in electronic contact networks has been studied by using mobile phone [10, 11, 136] and email networks [8, 202]. In addition, the spreading of computer viruses has been studied by looking at node activation times in email networks [203].

In Publication V and Publication VIII, we studied the theoretical maximum speed of information spreading in the mobile phone call network using an SI spreading model with the infection probability set to one. We found out that the temporal correlations in the call sequences considerable slow down spreading, and that taking the time stamps into account when modeling spreading is at least as important for the speed of the process as taking the topology of the social network into account. This is illustrated in Fig. 4.2. There are two main types of temporal correlations in communication networks that can affect the spreading speed: the bursty nature of link activation sequences [203, 204], and triggers and other correlations between neighboring nodes [8, 136]. Using temporal reference models we were able to show that most of the slowing down of

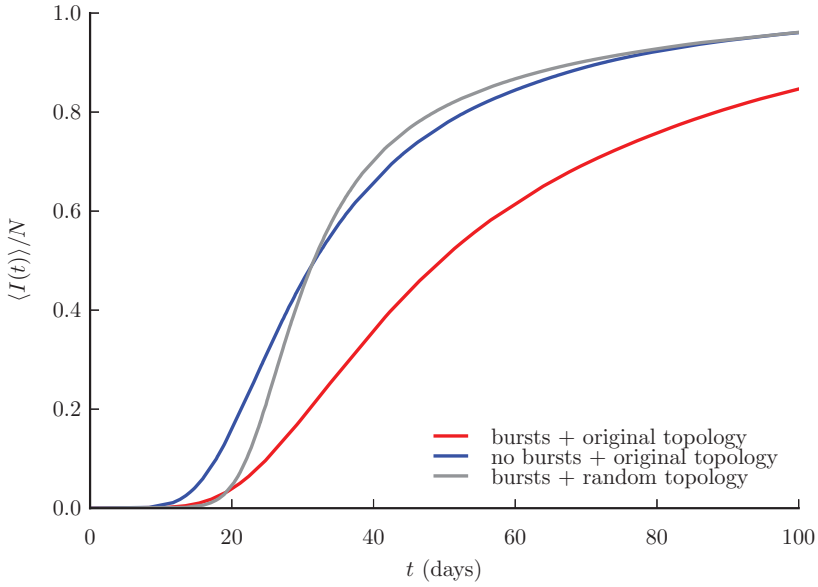


Figure 4.2. Spreading on the mobile phone network. The average number of infected people $\langle I \rangle$ as a function of time t, when the infection is started from a single node and susceptible nodes get infected every time they are in contact with the infected nodes. The averages are taken over 1000 realizations of the infection process. The red curve corresponds to the original temporal network containing both bursts and topology. The blue curve is from simulations on a reference network that retains the topology but destroys burstiness (time shuffled reference model). The reference network for the gray curve retains the bursty link sequences but the topology is randomized with the ER model (random network reference model). See Publication VIII for details.

spreading is due to the bursts of the link sequences, and that the triggers present in the call sequences have only minor consequences on the overall spreading speed. In Publication VIII we extensively studied the *relay times*, *i.e.* the times the infection waits at a node before it is transmitted across a link during the spreading process. We were able to analytically show a one-to-one correspondence between the burstiness measure and the expected relay time, and how the local correlations affect these times.

5. Summary of results and discussion

5.1 Summary

The publications in this Thesis can be divided to three categories depending on what types of networks they consider: Publications III, IV, and VI consider structure and dynamics of binary networks, such as network generation mechanisms, communities, and percolation, whereas Publications I, II, and VII concentrate on weighted networks, and Publications V and VIII deal with temporal contact networks.

5.1.1 Binary networks

In Publication III we categorize social network models and compare their ability to reproduce topological features of social networks. These models are based on network evolution, and mechanisms such as triadic closure and preferential attachment were found to produce fat-tailed degree distributions, decreasing clustering spectra and weakly assortative networks. However, some of the methods had problems in producing proper community structure. Further, the extent to which the structural properties match the real networks depends on the details of the microscopic network evolution mechanisms. Homophily-based models produce some unrealistic features, such as degree distributions that are too peaked and flat clustering spectra, but succeed in producing realistic assortativity. Thus, most of the important structural features of social networks are better explained by the microscopic mechanisms of network evolution models rather than homophily, but obviously the evolution of real social networks is controlled by combinations of both mechanisms.

Community structure in networks is one of the main research topics in complex networks. However, most of the work on the topic has concen-

trated on the issue of community detection, and not much attention has been given to the actual properties of the communities. In Publication IV we systematically study properties of communities in empirical social, information, communication, technological, and biological networks. We found that some community properties, such as broad size distributions, can be observed in all types of networks, and that there are many properties which are remarkably similar for networks of the same category.

In Publication VI we explored the possibility of finding applications for explosive percolation on real-world networks, since at the time the method had mostly been studied theoretically. We found out that the universality class of the percolation transition in explosive percolation depends on the type of topology the empirical network has and on the details of the irreversible percolation process. Further, we observed that the components near the critical threshold are related to the communities of the networks, both for empirical networks and for a simple model networks with built-in community structure.

5.1.2 Weighted networks

In Publication I generalizations of clustering coefficients to weighted networks were compared from a theoretical and empirical perspective. This comparative article was timely, since there were 4 different generalizations available at the time, and it was not clear what the differences between them are and which one should be used. The main conclusion was that the different generalizations could lead to completely different results even in very simple cases because of the different motivations behind the different coefficients. Also, the weighted clustering coefficient values were found to be highly redundant. This makes it hard to interpret the results since it is not clear if the coefficient values are mainly effected by the topology or the weight configuration. Thus, we suggest a different approach: one should first use the unweighted clustering coefficient to study the topology, and then study how weights are correlated with topology, using some other measures such as triangle intensities.

Publication II introduced the sequential clique percolation method, which is a fast method for detecting hierarchical clique percolation communities. The hierarchical approach was shown to reveal much more information about an online auctioning network than could have been possible by selecting a single threshold level in the hierarchy. The algorithm is also practical for normal clique percolation, and it was shown to be dramat-

ically faster for realistic model networks than the best alternative algorithm available at that point. Further, the worst-case scaling of the algorithm is almost linear in the number of k-cliques in the network.

In Publication VII the system formed from emotion concepts was studied for a first time as a weighted network. The standard way of studying emotion concepts is to project them into a metric space. We developed methods that allowed estimating when and how the assumption of an underlying metric space of emotions fails and the network perspective becomes useful. We further analyzed the emotion network with the SCP method developed in Publication II, which yielded an alternative point of view on relationships between emotion concepts.

5.1.3 Temporal networks

The exact timings of activations of edges in a complex social network formed from mobile phone calls of millions of subscribers and a smaller email network were studied in Publication V. The paper focused on the speed of spreading dynamics in such networks and considered how this is affected by different temporal aspects of the systems. It was previously known that community structure and Granovetter-type weight-topology correlations slow down spreading in social networks [119]. These effects were confirmed, but in addition we found a slowing-down effect caused by the time sequences. We were able to trace the effect down to the heterogeneity of the call sequences of individual links, and found out that daily patterns in the overall call frequency do not affect the spreading speed.

In Publication VIII, the results of Publication V were extended. More emphasis was given to the sequences of events on the links and to explaining how they slow down spreading. The effects of link sequences on the spreading speed were quantified by studying the relay times of the links both from theoretical and empirical perspectives. An analytical one-to-one correspondence between burstiness and the local spreading speed was found, and event triggering was shown to speed up spreading on links. Finally, the errors made by studying the link sequences in isolation instead of observing them as parts of the full network were quantified. We also formalized the reference model framework, making it easier to extend and to apply to new systems.

5.2 Discussion

The field of complex networks has come a long way since its initiation in the late 90's. In the beginning most of network science was concentrated on projecting all different types of complex systems to simple graphs and then applying general methods of analysis. This approach was extremely successful in the past, and it was sometimes even able to find universal properties of networks. However, in order to proceed and better understand specific systems, more information needs to be included than the pure topology of networks, and more specific questions about the networks need to be asked. Because of this, network science is becoming increasingly divided into specialized branches, and the methods are becoming more domain-specific and targeted *e.g.* mainly at social networks, intercellular networks, and brain-related networks. Thus, the use of weighted and temporal networks has become more common and will continue to do so in the future. This is also true for other special network types such as directed networks, interconnected networks, and spatially embedded networks. Although there has been a lot of pioneering work on weighted and temporal networks, the theory and tools for handling such networks have still not matured to the stage at which we are with binary networks.

As for any research area, it is important to begin with something as simple as possible, and only after the basics are fully understood, to continue to work on more advanced and complex topics and domain-specific problems. This has been one of the guidelines of this work. For example, when comparing the weighted clustering coefficients, we focused on identifying the underlying questions they aim at answering and on finding out what kinds of structure they find on small example networks, instead of directly jumping into applying them to empirical data. With the temporal contact networks, the starting point was the simplest possible dynamical process, which then provided an upper bound for the speed at which any dynamical process can proceed. Also, the temporal reference models were build such that more and more complicated correlations could be turned on in the networks, one by one, and with the relay times the starting point was to analytically solve the simplest case of uniformly random events on isolated edges before proceeding towards more complex phenomena. The author of this thesis believes that the systematic approach of explicitly defining the research questions the methods are build to answer while starting with the simplest possible questions is the only way to tackle

the complexities of weighted and temporal networks, and that the articles included in this Thesis have been an important contribution to this direction.

The research presented in this thesis opens up many doors for future research. For example, the online auction network of Publication II and the emotion concept network of Publication VII are not the only weighted networks with similar hierarchical overlapping community structure, which can now be extracted using the sequential clique percolation algorithm. In temporal communication networks, the impact of bursts and triggered events to spreading phenomenon are now better understood, but the exact way how they are related to topology, or attributes of the nodes such as age, sex, or location is still not known. Also, although other dynamic models on top of temporal networks have already been studied, all possible research lines related to using different temporal lattices and dynamic models are still far from being exhausted. Perhaps more importantly, some doors have also been closed, since there is no longer any justification for blindly selecting one of the weighted clustering coefficients without asking if this really is the type of weighted clustering that is relevant for this specific type of networks.

Bibliography

[1] M. Newman. Complex systems: A survey. *arXiv: 1112.1440* (2011).

[2] P. Anderson. More is different. *Science*, 177(4047):393 – 396 (1972).

[3] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442 (1998).

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512 (1999).

[5] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.*, 101(11):3747–3752 (2004).

[6] M. Granovetter. The strength of weak ties. *Am. J. Soc.*, 78:1360–1380 (1973).

[7] R. Albert, I. Albert, and G. L. Nakarado. Structural vulnerability of the north american power grid. *Phys. Rev. E*, 69:025103 (2004).

[8] P. Holme. Network reachability of real-world contact sequences. *Phys. Rev. E*, 71(4):046119 (2005).

[9] R. K. Pan and J. Saramäki. Path lengths, correlations, and centrality in temporal networks. *Phys. Rev. E*, 84:016105 (2011).

[10] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E*, 83(2):025102 (2011).

[11] M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, and M. Karsai. Multiscale analysis of spreading in a large communication network. *arXiv: 1112.4312* (2011).

[12] J. P. Scott. *Social Network Analysis: A Handbook*. Sage Publications Ltd (2000). ISBN 978-0761963394.

[13] P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290 (1959).

[14] B. Bollobás. *Random Graphs*. Cambridge University Press, second edition (2001). ISBN 0521797225.

[15] A. Rapoport. A study of a large sociogram. *Behavioral Science*, 6(4):279 (1961).

[16] S. Wasserman and F. Katherine. *Social Network Analysis: Methods and Applications*. Cambridge University Press (1994). ISBN 978-0521387071.

[17] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press (2003). ISBN 978-0198515906.

[18] S. Bornholdt and H. G. Schuster. *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH (2003). ISBN 978-3527403363.

[19] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press (2004). ISBN 978-0521826983.

[20] M. Newman. *Networks: An Introduction*. Oxford University Press (2010). ISBN 978-0199206650.

[21] S. N. Dorogovtsev. *Lectures on Complex Networks*. Oxford University Press (2010).

[22] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97 (2002).

[23] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):pp. 167–256 (2003).

[24] S. Boccaletti, V. Latora, Y. Moreno, Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Phys. Rep.*, 424:175 —- 308 (2006).

[25] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Critical phenomena in complex networks. *Rev. Mod. Phys.*, 80:1275–1335 (2008).

[26] A. L. Barabási. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume (2003). ISBN 978-0452284395.

[27] D. J. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company (2004). ISBN 978-0393325423.

[28] F. Kopos. *Biological Networks*. World Scientific Publishing Company (2007). ISBN 978-9812706959.

[29] B. H. Junker and F. Schreiber. *Analysis of Biological Networks*. Wiley-Interscience (2008). ISBN 978-0470041444.

[30] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press (2008). ISBN 978-0521879507.

[31] R. Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, Heidelberg, fourth edition (2010).

[32] R. Albert, H. Jeong, and A.-L. Barabási. Internet: Diameter of the world-wide web. *Nature*, 401:130–131 (1999).

[33] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Netw.*, 33:309–320 (2000).

[34] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29:251–262 (1999).

[35] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913 (2002).

[36] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and B. A.-L. The large-scale organization of metabolic networks. *Nature*, 407:651–655 (2000).

[37] H. Ma and A.-P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277 (2003).

[38] R. Khanin and E. Wit. How scale-free are biological networks. *Journal of Computational Biology*, 13(3):810–818 (2006).

[39] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703 (2009).

[40] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. U.S.A.*, 97(21):11149–11152 (2000).

[41] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203 (2001).

[42] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382 (2000).

[43] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701 (2002).

[44] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126 (2003).

[45] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701 (2001).

[46] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112 (2003).

[47] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Phys. Rev. E*, 65:066122 (2002).

[48] S. Milgram. The small world problem. *Psychology Today*, 1 (1967).

[49] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443 (1969).

[50] D. Stauffer and A. Aharony. *Introduction to percolation theory*. Taylor & Francis (1994). ISBN 9780748402533.

[51] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85:5468–5471 (2000).

[52] C. Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Phys. Rev. E*, 61:5678–5682 (2000).

[53] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118 (2001).

[54] R. Cohen, D. ben Avraham, and S. Havlin. Percolation critical exponents in scale-free networks. *Phys. Rev. E*, 66:036113 (2002).

[55] M. E. J. Newman. Component sizes in networks with arbitrary degree distributions. *Phys. Rev. E*, 76:045101 (2007).

[56] D. Achlioptas, R. M. D'Souza, and J. Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455 (2009).

[57] N. A. M. Araújo and H. J. Herrmann. Explosive percolation via control of the largest cluster. *Phys. Rev. Lett.*, 105:035701 (2010).

[58] E. J. Friedman and A. S. Landsberg. Construction and analysis of random networks with explosive percolation. *Phys. Rev. Lett.*, 103:255701 (2009).

[59] R. A. da Costa, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Explosive percolation transition is actually continuous. *Phys. Rev. Lett.*, 105:255701 (2010).

[60] J. Nagler, A. Levina, and M. Timme. Impact of single links in competitive percolation. *Nature Physics*, 7:265–270 (2011).

[61] R. M. Ziff. Scaling behavior of explosive percolation on the square lattice. *Phys. Rev. E*, 82:051105 (2010).

[62] F. Radicchi and S. Fortunato. Explosive percolation in scale-free networks. *Phys. Rev. Lett.*, 103:168701 (2009).

[63] Y. S. Cho, J. S. Kim, J. Park, B. Kahng, and D. Kim. Percolation transitions in scale-free networks under the achlioptas process. *Phys. Rev. Lett.*, 103:135702 (2009).

[64] H. D. Rozenfeld, L. K. Gallos, and H. A. Makse. Explosive percolation in the human protein homology network. *Eur. Phys. J. B*, 75(3):305–310 (2010).

[65] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70:056131 (2004).

[66] R. K. Pan and S. Sinha. Modularity produces small-world networks with dynamical time-scale separation. *Europhys. Lett.*, 85(6):68006 (2009).

[67] M. Kivelä. *A network perspetive on the genetic population structure of seagrass Posidonia oceanica*. Master's thesis, Helsinki University of Technology (2009).

[68] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35:66–71 (2002).

[69] S. Fortunato. Community detection in graphs. *Phys. Reports*, 486:75–174 (2010).

[70] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814—-818 (2005).

[71] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764 (2010).

[72] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117 (2009).

[73] J. Kleinberg. An impossibility theorem for clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 446–453. MIT Press (2002).

[74] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826 (2002).

[75] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78:046110 (2008).

[76] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80:016118 (2009).

[77] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015 (2009).

[78] H. Shen, X. Cheng, K. Cai, and M.-B. Hu. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706 – 1712 (2009).

[79] T. S. Evans. Clique graphs and overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(12):P12037 (2010).

[80] Y. Kim and H. Jeong. Map equation for link communities. *Phys. Rev. E*, 84:026110 (2011).

[81] S. Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02017 (2011).

[82] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113 (2004).

[83] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133 (2004).

[84] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111 (2004).

[85] A. Gog, D. Dumitrescu, and B. Hirsbrunner. Community detection in complex networks using collaborative evolutionary algorithms. In *Advances in Artificial Life*, volume 4648 of *Lecture Notes in Computer Science*, pages 886–894. Springer Berlin / Heidelberg (2007). ISBN 978-3-540-74912-7.

[86] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104 (2006).

[87] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8577–8582 (2006).

[88] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101 (2004).

[89] G. Agarwal and D. Kempe. Modularity-maximizing graph communities via mathematical programming. *Eur. Phys. J. B*, 66(3):409–418 (2008).

[90] T. N. Dinh and M. T. Thai. Finding community structure with performance guarantees in complex networks. In *Proceedings of the 3rd IEEE International Conference on Social Computing*, SOCIALCOM (2011).

[91] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2):172 –188 (2008).

[92] B. H. Good, Y.-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106 (2010).

[93] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A.*, 104(1):36–41 (2007).

[94] J. Reichardt and S. Bornholdt. Partitioning and modularity of graphs with arbitrary degree distribution. *Phys. Rev. E*, 76:015102 (2007).

[95] M. Rosvall and C. T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104(18):7327–7331 (2007).

[96] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.*, 105(4):1118–1123 (2008).

[97] A. Viamontes Esquivel and M. Rosvall. Compression of flow can reveal overlapping-module organization in networks. *Phys. Rev. X*, 1:021025 (2011).

[98] M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS ONE*, 6(4):e18209 (2011).

[99] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630 (1957).

[100] J. Park and M. E. J. Newman. Statistical mechanics of networks. *Phys. Rev. E*, 70:066117 (2004).

[101] T. Squartini and D. Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8):083001 (2011).

[102] P. Erdős and A. Rényi. On the Evolution of Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61 (1960).

[103] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biophysics*, 13 (1951).

[104] E. N. Gilbert. Enumeration of labelled graphs. *Canad. J. Math.*, 8:405 – 411 (1956).

[105] A. Békéssy, P. Bekessy, and J. Komlós. Asymptotic enumeration of regular matrices. *Studia Scientiarum Mathematicarum Hungarica*, 7:343–353 (1972).

[106] E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307 (1978).

[107] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2-3):161–179 (1995).

[108] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.*, 7:295 (1998).

[109] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842 (1986).

[110] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61(3):401–425 (1996).

[111] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153 (2006).

[112] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191 (2007).

[113] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192 – 215 (2007).

[114] C. J. Geyer and E. A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699 (1992).

[115] T. A. B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3 (2002).

[116] D. D. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306 (1976).

[117] J. R. Banavar, A. Maritan, and A. Rinaldo. Size and form in efficient transportation networks. *Nature*, 399:130 – 132 (1999).

[118] W. Li and X. Cai. Statistical analysis of airport network of china. *Phys. Rev. E*, 69:046106 (2004).

[119] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104(18):7332–7336 (2007).

[120] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.*, 9(6):179 (2007).

[121] K. S. Gleditsch. Expanded trade and gdp data. *Journal of Conflict Resolution*, 46:712–24 (2002).

[122] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev. E*, 75:027105 (2007).

[123] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64(1):016131+ (2001).

[124] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132+ (2001).

[125] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.*, 98(2):404–409 (2001).

[126] Y. Iturria-Medina, E. Canales-Rodríguez, L. Melie-García, P. Valdés-Hernández, E. Martínez-Montes, Y. Alemán-Gómez, and J. Sánchez-Bornot. Characterizing brain anatomical connections using diffusion weighted mri and graph theory. *NeuroImage*, 36(3):645–660 (2007).

[127] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS Biol*, 6(7):e159 (2008).

[128] C.-Y. Lo, P.-N. Wang, K.-H. Chou, J. Wang, Y. He, and C.-P. Lin. Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in alzheimer's disease. *The Journal of Neuroscience*, 30(50):16876–16885 (2010).

[129] J.-P. Onnela, K. Kaski, and J. Kertész. Clustering and information in correlation based financial networks. *Eur. Phys. J. B*, 38:353 – 362 (2004).

[130] A. F. Rozenfeld, S. Arnaud-Haond, E. Hernández-García, V. M. Eguíluz, E. A. Serrão, and C. M. Duarte. Network analysis identifies weak and strong links in a metapopulation system. *Proc. Natl. Acad. Sci. U.S.A.*, 105(48):18824–18829 (2008).

[131] E. Hernández-García, A. F. Rozenfeld, V. M. Eguíluz, S. Arnaud-Haond, and C. M. Duarte. Clone size distributions in networks of genetic similarity. *Physica D: Nonlinear Phenomena*, 224(1–2):166 – 173 (2006).

[132] R. Toivonen, M. Kivelä, J. Saramäki, M. Viinikainen, M. Vanhatalo, and M. Sams. Networks of emotion concepts. *PLoS ONE*, 7(1):e28883 (2012).

[133] E. Almaas, B. Kovács, V. T., O. Z. N., and A.-L. Barabási. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427:839 – 843 (2004).

[134] A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor. Compartments revealed in food-web structure. *Nature*, 426:282 – 285 (2003).

[135] S. E. Ahnert, D. Garlaschelli, T. M. A. Fink, and G. Caldarelli. Ensemble approach to the analysis of weighted networks. *Phys. Rev. E*, 76:016101 (2007).

[136] G. Miritello, E. Moro, and R. Lara. The dynamical strength of social ties in information spreading. *Phys. Rev. E*, 83:045102 (2011).

[137] P. Holme, F. Liljeros, C. R. Edling, and B. J. Kim. Network bipartivity. *Phys. Rev. E*, 68:056107 (2003).

[138] T. Zhou, J. Ren, M. c. v. Medo, and Y.-C. Zhang. Bipartite network projection and personal recommendation. *Phys. Rev. E*, 76:046115 (2007).

[139] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680 (1946).

[140] J. Gait. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87(3):564–567 (1980).

[141] P. F. Velleman and L. Wilkinson. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1):65–72 (1993).

[142] P. Grindrod. Range-dependent random graphs and their application to modeling large small-world proteome datasets. *Phys. Rev. E*, 66:066702 (2002).

[143] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):17 (2005).

[144] P. Holme, S. M. Park, B. J. Kim, and C. R. Edling. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A: Statistical Mechanics and its Applications*, 373(0):821 – 830 (2007).

[145] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E*, 71(6):065103 (2005).

[146] M. Serrano, M. Boguñá, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.*, 106(16):6483–6488 (2009).

[147] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, second edition (2001).

[148] G. Palla, I. Derényi, and T. Vicsek. The critical point of k-clique percolation in the Erdős–Rényi graph. *Journal of Statistical Physics*, 128:219–227 (2007).

[149] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Phys. Rev. Lett.*, 94:160202 (2005).

[150] B. Bollobás and O. Riordan. Clique percolation. *Random Structures & Algorithms*, 35(3):294–322 (2009).

[151] I. Farkas, D. Ábel, G. Palla, and T. Vicsek. Weighted network modules. *New Journal of Physics*, 9(6):180+ (2007).

[152] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and V. T. Cfinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22:1021–1023 (2006).

[153] R. Karp. Reducibility among combinatorial problems. In R. Miller and J. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press (1972).

[154] C. Bron and J. Kerbosch. Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577 (1973).

[155] E. Tomita, A. Tanaka, and H. Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42 (2006).

[156] D. Eppstein and D. Strash. Listing all maximal cliques in large sparse real-world graphs. In P. Pardalos and S. Rebennack, editors, *Experimental Algorithms*, volume 6630 of *Lecture Notes in Computer Science*, pages 364–375. Springer Berlin / Heidelberg (2011).

[157] M. Valencia, J. Martinerie, S. Dupont, and M. Chavez. Dynamic small-world behavior in functional brain networks unveiled by an event-related networks approach. *Phys. Rev. E*, 77:050905 (2008).

[158] S. I. Dimitriadis, N. A. Laskaris, V. Tsirka, M. Vourkas, S. Micheloyannis, and S. Fotopoulos. Tracking brain dynamics via time-dependent network analysis. *Journal of Neuroscience Methods*, 193(1):145–155 (2010).

[159] D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. U.S.A.*, 108(18):7641–7646 (2011).

[160] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93 (2004).

[161] K. Komurov and M. White. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol*, 3 (2007).

[162] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27(2):199–204 (2009).

[163] T. M. Przytycka, M. Singh, and D. K. Slonim. Toward the dynamic interactome: it's about time. *Briefings in Bioinformatics*, 11(1):15–29 (2010).

[164] R. Yoshida, S. Imoto, and T. Higuchi. Estimating time-dependent gene networks from time series microarray data by dynamic linear models with markov switching. In *Proc IEEE Comput Syst Bioinform Conf.*, CSB '05, pages 289–298. IEEE Computer Society (2005).

[165] A. Rao, A. Hero, D. States, and J. Engel. Inferring time-varying network topologies from gene expression data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007(1):51947 (2007).

[166] S. Lèbre, J. Becq, F. Devaux, M. Stumpf, and G. Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(1):130 (2010).

[167] J. P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész. Dynamic asset trees and portfolio analysis. *Eur. Phys. J. B*, 30:285–288 (2002).

[168] J. P. Onnela, A. Chakraborti, K. Kaska, and J. Kertész. Dynamic asset trees and black monday. *Physica A*, 324:247–252 (2003).

[169] J. P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto. Asset trees and asset graphs in financial markets. *Physica Scripta*, T106:48–54 (2003).

[170] A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro. Time-varying graphs and dynamic networks. In *Proceedings of the 10th international conference on Ad-hoc, mobile, and wireless networks*, ADHOC-NOW'11, pages 346–359. Springer-Verlag, Berlin, Heidelberg (2011). ISBN 978-3-642-22449-2.

[171] J. Tang, S. Scellato, M. Musolesi, C. Mascolo, and V. Latora. Small-world behavior in time-varying graphs. *Phys. Rev. E*, 81:055101 (2010).

[172] B. B. Xuan, A. Ferreira, and A. Jarry. Computing shortest, fastest, and foremost journeys in dynamic networks. *International Journal of Foundations of Computer Science*, 14(2):267–285 (2003).

[173] D. Kempe. Connectivity and inference problems for temporal networks. *Journal of Computer and System Sciences*, 64(4):820–842 (2002).

[174] P. Holme and J. Saramäki. Temporal networks. *arXiv: 1108.1780* (2011).

[175] A. L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211 (2005).

[176] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A: Math. Theor.*, 41(22):224015 (2008).

[177] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *Europhys. Lett.*, 81:48002 (2008).

[178] L. Kovanen. *Structure and dynamics of a large-scale complex social network*. Master's thesis, Helsinki University of Technology (2009).

[179] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005 (2011).

[180] D. H. Zanette. Dynamics of rumor propagation on small-world networks. *Phys. Rev. E*, 65:041908 (2002).

[181] Y. Moreno, M. Nekovee, and A. F. Pacheco. Dynamics of rumor spreading in complex networks. *Phys. Rev. E*, 69:066130 (2004).

[182] V. M. Eguíluz and M. G. Zimmermann. Transmission of information and herd behavior: An application to financial markets. *Phys. Rev. Lett.*, 85:5659–5662 (2000).

[183] A. T. Bernardes, D. Stauffer, and J. Kertész. Election results and the sznajd model on barabasi network. *Eur. Phys. J. B*, 25(1):123–127 (2002).

[184] K. Suchecki, V. M. Eguíluz, and M. San Miguel. Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution. *Phys. Rev. E*, 72:036132 (2005).

[185] V. Sood and S. Redner. Voter model on heterogeneous graphs. *Phys. Rev. Lett.*, 94:178701 (2005).

[186] A. G. Smart, L. A. N. Amaral, and J. M. Ottino. Cascading failure and robustness in metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.*, 105(36):13223–13228 (2008).

[187] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025–1028 (2010).

[188] R. Anderson and M. R.M. *Infectious Diseases of Humans: Dynamics and Control*. Oxford Science Publications (1992).

[189] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60:7332–7342 (1999).

[190] D. Mollison. Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(3):283–326 (1977).

[191] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157 – 172 (1982).

[192] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Phys. Rev. E*, 64:066112 (2001).

[193] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 26:521–529 (2002).

[194] R. Pastor-Satorras and A. Vespignani. Immunization of complex networks. *Phys. Rev. E*, 65:036104 (2002).

[195] R. Cohen, S. Havlin, and D. ben Avraham. Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.*, 91:247901 (2003).

[196] J. Stehle, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Regis, J.-F. Pinton, N. Khanafer, W. Van den Broeck, and P. Vanhems. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 9(1):87 (2011).

[197] M. Morris. Concurrent partnerships and transmission dynamics in networks. *Social Networks*, 17(3-4):299–318 (1995).

[198] M. Morris and M. Kretzschmar. Concurrent partnerships and the spread of hiv. *AIDS*, 11(5):641—648 (1997).

[199] L. E. C. Rocha, F. Liljeros, and P. Holme. Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proceedings of the National Academy of Sciences*, 107(13):5706–5711 (2010).

[200] L. E. C. Rocha, F. Liljeros, and P. Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*, 7(3):e1001109 (2011).

[201] F. Liljeros, J. Giesecke, and P. Holme. The contact network of inpatients in a regional healthcare system. a longitudinal case study. *Mathematical Population Studies*, 14(4):269–284 (2007).

[202] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 435–443. ACM (2008).

[203] A. Vazquez, B. Rácz, A. Lukács, and A.-L. Barabási. Impact of non-poissonian activity patterns on spreading processes. *Phys. Rev. Lett.*, 98(15):158702 (2007).

[204] B. Min, K.-I. Goh, and A. Vazquez. Spreading dynamics following bursty human activity patterns. *Phys. Rev. E*, 83(3):036102 (2011).

Aalto-DD 103/2012

ISBN 978-952-60-4758-4
ISBN 978-952-60-4759-1 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
Aalto University School of Science
Department of Biomedical Engineering and Computational Science

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS